Segment-Level Pyramid Match Kernels For The Classification of Varying Length Patterns of Speech Using SVMs

Shikha Gupta and A. D. Dileep
School of Computing and Electrical Engineering
Indian Institute of Technology Mandi
Mandi, H.P., India

Veena Thenkanidiyoor
Department of Computer Science and Engineering
National Institute of Technology Goa
Ponda, Goa, India

***Abstract*—Classification of long duration speech, represented as varying length sets of feature vectors using support vector machine (SVM) requires a suitable kernel. In this paper we propose a novel segment-level pyramid match kernel (SLPMK) for the classification of varying length patterns of long duration speech represented as sets of feature vectors. This kernel is designed by partitioning the speech signal into increasingly finer segments and matching the corresponding segments. We study the performance of the SVM-based classifiers using the proposed SLPMKs for speech emotion recognition and speaker identification and compare with that of the SVM-based classifiers using other dynamic kernels.**

I. Introduction

A speech utterance is subjected to short-time analysis that involves performing spectral analysis on each frame of about 20 milliseconds duration. Each frame of speech is represented by a real valued feature vector. The duration of the utterances varies from one utterance to another. Hence, the number of frames also differs from one utterance to another. The speech signal of an utterance with T frames is represented as a sequential pattern $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t$ is a feature vector for frame t. In the tasks such as speaker identification, spoken language identification, and speech emotion recognition. The duration of data is long and preserving sequence information is not critical. Hence a speech signal is represented as set of feature vectors. The focus of this paper is on classification of varying length patterns of long duration speech that are represented as sets of continuous valued feature vectors. Conventionally, Gaussian mixture models (GMMs) [1] are used for classification of varying length patterns represented as sets of feature vectors. The maximum likelihood (ML) based method is commonly used for estimation of parameters of the GMM for each class. When the amount of the training data available per class is limited, robust estimates of model parameters can be obtained through maximum a posteriori adaptation of the class-independent GMM (CIGMM), which is also called as universal background model (UBM), to the training data of each class [2]. The CIGMM or UBM is a large GMM trained using the training data of all the classes. Classification of varying length sets of feature vectors using SVM-based classifiers requires the design of a suitable kernel as a measure of similarity between a pair of sets of feature vectors. The kernels designed for varying length patterns are referred to as dynamic kernels [3]. Fisher kernel using GMM-based likelihood score vectors [4], probabilistic sequence kernel [5], GMM supervector kernel [6], GMM-UBM mean interval kernel [7], GMM-based intermediate matching kernel [3] and GMM-based pyramid match kernel [8] are some of the state-of-the-art dynamic kernels for sets of feature vectors.

In this paper, we propose segment-level pyramid match kernel (SLPMK) for SVM-based classification of speech signals represented as varying length sets of feature vectors. We propose to repeatedly divide a speech signal to form a pyramid of increasingly finer segments. Then the SLPMK between a pair of speech signals is constructed by matching the corresponding segments at every level of the pyramid. We propose two approaches to obtain SLPMK. The first approach is inspired by the spatial pyramid match kernel [9] proposed for image classification. In this approach, each segment is represented as a bag-of-codewords, where the codewords are obtained by clustering all the feature vectors of all the speech signals using K-means clustering technique. The codebook-based SLPMK (CBSLPMK) between a pair of speech signals is computed as a weighted sum of the number of new matches found at different levels of the pyramid of segments. The bag-of-codewords representation used in CBSLPMK suffers from loss of information due to the hard assignment of a feature vector to a codeword. To address this issue, we propose Gaussian mixture model (GMM) based SLPMK (GMMSLPMK) as second approach to construct the SLPMK. In this approach, bag-of-codewords representation for each segment of a speech signal is obtained by soft assignment of the feature vectors to codewords using class independent GMM as soft clustering technique. Salient features of the proposed SLPMK are: (i) maintaining the temporal ordering of the feature vectors in a speech signal for some extent, and (ii) using the local information for matching between two speech utterances represented as sets of feature vectors.

In Section II, a review of dynamic kernels for sets of feature vectors is presented. The proposed SLPMKs for sets of feature

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

vectors is described in Section III. In Section IV we present our studies. The conclusion is presented in Section V.

II. Dynamic kernels for sets of feature vectors

In this section, we review the approaches to design dynamic kernels for varying length patterns represented as sets of feature vectors. The Fisher kernel (FK) using GMM-based likelihood score vectors [4] and the probabilistic sequence kernel (PSK) [5], are the dynamic kernels for sets of feature vectors constructed using the explicit mapping based approaches. The FK uses a GMM for mapping a set of feature vectors onto a Fisher score-space. The Fisher score-space for a class is obtained using the first order derivatives of the log likelihood output of GMM for that class with respect to the GMM parameters. The PSK maps a set of feature vectors onto a high dimensional probabilistic score space. The probabilistic score space for a class is obtained using the posterior probability of components of the GMM built for that class.

The GMM supervector kernel (GMMSVK) [6] and the GMM-UBM mean interval kernel (GUMIK) [7], are designed using the probabilistic distance metric based approaches. The GMMSVK uses example-specific adapted GMM built for each example by adapting the mean vectors of the UBM using the data of that example. The GMMSVK is then computed between the pair of examples by computing the KL-divergence between the pair of example-specific adapted GMMs. The GUMIK uses example-specific adapted GMM built for each example by adapting the mean vectors and covariance matrices of the UBM using the data of that example. The GUMIK is then computed between the pair of examples by computing the Bhattacharyya distance between the pair of example-specific adapted GMMs.

GMM-based intermediate matching kernel (GMMIMK) [3] and GMM-based pyramid match kernel (GMMPMK) [8] are the dynamic kernels designed using the matching based approach. An intermediate matching kernel (IMK) [10] is constructed by matching the sets of feature vectors using a set of virtual feature vectors. For every virtual feature vector, a feature vector is selected from each set of feature vectors and a base kernel (such as Gaussian kernel) for the two selected feature vectors is computed. The IMK for a pair of sets of feature vectors is computed as a combination of these base kernels. In [3], the set of virtual feature vectors considered are in the form of the components of class independent GMM (CIGMM). For every component of CIGMM, a feature vector each from the two sets of feature vectors, which has the highest probability of belonging to that component (i.e., value of responsibility term) is selected and a base kernel is computed between the selected feature vectors. In the PMK [11], a set of feature vectors is mapped onto a multi-resolution histogram pyramid. The kernel is computed between a pair of examples by matching the pyramids using a weighted histogram intersection match function at each level of pyramid. In [8], the CIGMMs built with increasingly larger number of components are used to construct the histograms at the different levels in the pyramid. In our studies, we compare the performance of the proposed SLPMK-based SVM classifier with the performance of SVM-based classifiers using kernels reviewed in this section.

III. Segment-level Pyramid Match Kernels

In designing SLPMK, a speech utterance represented as a set of feature vectors is decomposed into pyramid of increasingly finer segments. SLPMK between a pair of speech utterances is computed by matching the corresponding segments at each level in the pyramid. Let $j = 0, 1, \ldots, J-1$ be the J levels in the pyramid. At level 0 (i.e. j=0) complete speech signal is considered as a segment. At level 1 (i.e. j=1), a speech signal is divided into two equal segments. At level 2 (i.e. j=2), a speech signal is divided into four equal segments and so on. Hence at any level j, a speech utterance is partitioned into 2^j equal segments.

A. Codebook based SLPMK

For designing codebook based segment-level pyramid match kernel (CBSLPMK) we borrowed the idea from spatial pyramid match kernel [9] which consider the pyramid of spatial division of images. In designing CBSLPMK, every segment from a speech utterance is mapped to a bag-of-codewords representation. A codeword is a representative feature vector for a group of similar feature vectors. Collection of all the codewords is known as a codebook. A codebook of size Q is constructed by clustering the feature vectors in the training examples of all the classes using K-means clustering technique. The bag-of-codewords representation for a speech segment is obtained by assigning every feature vector to one of the Q codewords. Let $\mathbf{h}_{jk}(\mathbf{X})$ be the Q-dimensional bag-of-codewords representation of kth segment of an example $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ in the jth level of pyramid. Let $h_{jkq}(\mathbf{X})$ be an element in the $\mathbf{h}_{jk}(\mathbf{X})$, indicating the number of feature vectors of kth segment assigned to qth codeword. Let $\mathbf{X}_m = \{\mathbf{x}_{m1}, \mathbf{x}_{m2}, ..., \mathbf{x}_{mT_m}\}$ and $\mathbf{X}_n = \{\mathbf{x}_{n1}, \mathbf{x}_{n2}, ..., \mathbf{x}_{nT_n}\}$ be the two sets of feature vectors. The number of matches in the qth codeword between the kth segments of $\mathbf{X}_m$ and $\mathbf{X}_n$ at jth level of pyramid is given by

$$s_{jkq} = \min(h_{jkq}(\mathbf{X}_m), h_{jkq}(\mathbf{X}_n)) \tag{1}$$

Total number of matches at level j between the kth segments of $\mathbf{X}_m$ and $\mathbf{X}_n$ is obtained as,

$$S_{jk} = \sum_{q=1}^{Q} s_{jkq} \tag{2}$$

Total number of matches between $\mathbf{X}_m$ and $\mathbf{X}_n$ at level j is obtained as,

$$\hat{S}_j = \sum_{k=1}^{2^j} S_{jk} \tag{3}$$

Note that the number of matches found at level j also includes all the matches found at the finer level $j+1$. Therefore, the

number of new matches found at level j is given by $\hat{S}_j$ - $\hat{S}_{j+1}$. The CBSLPMK is computed as,

$$K_{\text{CBSLPMK}}(\mathbf{X}_m, \mathbf{X}_n) = \sum_{j=0}^{J-2} \frac{1}{2^{J-(j+1)}} (\hat{S}_j - \hat{S}_{j+1}) + \hat{S}_{J-1} \quad (4)$$

The key issue in the design of SLPMK is the choice of the technique for constructing the bag-of-codewords representation for each segment of speech utterance. The K-means clustering method makes use of information about the centers of clusters and the distances of a feature vector to the centers of clusters to assign that feature vector to one of the clusters. A better bag-of-codewords representation of speech segment can be obtained by considering a clustering method that considers additional information like the spread of the clusters and the sizes of the clusters along with the centers of the clusters [11]. Moreover, the construction of CBSLPMK involves hard clustering. A better SLPMK is constructed by using soft clustering. In the next subsection, we propose the GMM-based SLPMK. The GMM uses the information about the spread and the size of the clusters along with the centers of the clusters for soft assignment of feature vectors.

B. GMM-based SLPMK

In this approach, we propose to use a class-independent GMM (CIGMM) for forming the clusters to obtain the bag-of-codewords representation for each speech segment. CIGMM is a large GMM of Q components built using the feature vectors in the training examples of all the classes. Every component of the CIGMM represents a codeword. The q^{th} codeword is now represented by the mean vector $\boldsymbol{\mu}_q$, covariance matrix $\boldsymbol{\Sigma}_q$ and mixture weight ω_q of the qth component of CIGMM. The soft assignment of a feature vector from a segment to the qth component in the CIGMM is obtained using the responsibility term and it is given by

$$\gamma_q(\mathbf{x}_t) = \frac{w_q \boldsymbol{\mathcal{N}}(\mathbf{x}_t | \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)}{\sum_{q'=1}^{Q} w_{q'} \boldsymbol{\mathcal{N}}(\mathbf{x}_t | \boldsymbol{\mu}_{q'}, \boldsymbol{\Sigma}_{q'})} \quad (5)$$

where $\boldsymbol{\mathcal{N}}(\mathbf{x}_t \mid \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ is the normal density for the component q. For the kth speech segment at jth level of pyramid, the effective number of feature vectors $h_{jkq}(\mathbf{X})$ assigned to a component q is given by

$$h_{jkq}(\mathbf{X}) = \sum_{t=1}^{T_k} \boldsymbol{\gamma}_q(\mathbf{x}_t) \quad (6)$$

where, T_k is the number of feature vectors in the kth segment of $\mathbf{X}$. For a pair of examples represented as sets of feature vectors, $\mathbf{X}_m$ and $\mathbf{X}_n$, number of matches in the qth codeword between the kth segments of $\mathbf{X}_m$ and $\mathbf{X}_n$ at jth level of pyramid (s_{jkq}), total number of matches at level j between the kth segments (S_{jk}) and total number of matches between $\mathbf{X}_m$ and $\mathbf{X}_n$ at level j ($\hat{S}_j$) are computed as in (1), (2) and (3) respectively. The GMM-based SLPMK (GMMSLPMK) between a pair of examples $\mathbf{X}_m$ and $\mathbf{X}_n$, K_{GMMSLPMK} is then computed as in (4).

Both CBSLPMK and GMMSLPMK are valid positive definite kernel. The proof for the CBSLPMK and GMMSLPMK as positive definite kernel is excluded due to the limitation of pages. The main advantages of using SLPMK over other dynamic kernels, especially over GMMPMK [8] are: (i) use of local information while matching a pair of speech utterances and (ii) maintaining temporal ordering of feature vectors in a speech utterance for some extent by matching at segment levels.

IV. Experimental studies

In this section, effectiveness of the proposed kernels is studied for speech emotion recognition and speaker identification tasks using SVM-based classifiers. We have considered Mel frequency cepstral coefficients (MFCC) as features. A frame size of 20 ms and a shift of 10 ms are used for feature extraction from the speech signal of an utterance. Every frame is represented using a 39-dimensional feature vector. Here, the first 12 features are Mel frequency cepstral coefficients and the 13th feature is log energy. The remaining 26 features are the delta and acceleration coefficients. We consider, LIBSVM [12] tool to build the SVM-based classifiers. In this study, one-against-the-rest approach is considered for 7-class and 4-class speech emotion recognition tasks and 122-class speaker identification task. The value of trade-off parameter, C in SVM is chosen empirically as 10^{-3}.

The Berlin emotional speech database (Emo-DB) [13] and the German FAU Aibo emotion corpus (FAU-AEC) [14] are used for studies on speech emotion recognition task. Emo-DB contains 494 utterances belonging to seven emotional categories. The multi-speaker speech emotion recognition accuracy presented is the average classification accuracy along with 95% confidence interval obtained for 5-fold stratified cross-validation. In FAU-AEC, we have considered four super classes of emotions such as 'anger', 'emphatic', 'neutral', and 'motherese'. We have considered an almost balanced subset of the corpus defined for these four classes by CEICES of the Network of Excellence HUMAINE funded by European Union [14]. We perform the classification at the chunk (speech utterance) level in the Aibo chunk set. The speaker-independent speech emotion recognition accuracy presented is the average classification accuracy along with 95% confidence interval obtained for 3-fold stratified cross validation. The 3-fold cross validation is based on the three splits defined in Appendix A.2.10 of [14].

Speaker identification experiments are performed on the 2002 and 2003 NIST speaker recognition (SRE) corpora [15], [16]. We have considered the 122 male speakers that are common to the 2002 and 2003 NIST SRE corpora. Training data for a speaker includes a total of about 3 minutes of speech from the single conversations in the training set of 2002 and 2003 NIST SRE corpora. The test data from the 2003 NIST SRE corpus is used for testing the speaker recognition systems. The speaker identification accuracy presented is the classification accuracy obtained for test examples. The training

TABLE I
CLASSIFICATION ACCURACY (CA) (IN %) OF THE SVM-BASED CLASSIFIERS WITH CBSLPMK AND GMMSLPMK FOR SPEECH EMOTION RECOGNITION (SER) AND SPEAKER IDENTIFICATION (SPK-ID) TASKS FOR THE DIFFERENT VALUES OF Q AND J. HERE, CA95%CI INDICATES AVERAGE CLASSIFICATION ACCURACY ALONG WITH 95% CONFIDENCE INTERVAL.

Q	J	SVM using CBSLPMK			SVM using GMMSLPMK		
		SER		Spk-ID	SER		Spk-ID
		Emo-DB	FAU-AEC		Emo-DB	FAU-AEC	
		CA95%CI	CA95%CI	(CA)	CA95%CI	CA95%CI	(CA)
256	1	79.80±0.15	59.09±0.06	77.82	87.20±0.20	64.79±0.10	79.50
	2	81.20±0.19	59.73±0.10	78.09	88.00±0.17	64.99±0.06	80.94
	3	80.00±0.20	59.23±0.09	77.65	87.80±0.19	64.75±0.05	79.84
512	1	82.80±0.20	57.45±0.08	79.54	88.20±0.15	65.27±0.08	81.08
	2	**87.60±0.20**	60.32±0.07	80.67	90.80±0.17	66.03±0.07	84.79
	3	84.60±0.18	59.98±0.09	79.43	**92.40±0.19**	65.43±0.1	83.51
1024	1	82.40±0.16	59.75±0.10	81.54	88.80±0.16	66.93±0.09	87.18
	2	84.20±0.19	60.88±0.05	**84.85**	88.60±0.17	**67.97±0.10**	**91.35**
	3	83.20±0.17	**61.04±0.09**	83.34	89.60±0.17	67.45±0.09	89.23

and test datasets as defined in the NIST SRE corpora are used in studies.

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY (CA) (IN %) OF THE GMM-BASED CLASSIFIERS AND SVM-BASED CLASSIFIERS USING STATE-OF-THE-ART DYNAMIC KERNELS (MENTIONED IN SECTION II) FOR SPEECH EMOTION RECOGNITION (SER) TASK AND SPEAKER IDENTIFICATION (SPK-ID) TASK. HERE, CA95%CI INDICATES AVERAGE CLASSIFICATION ACCURACY ALONG WITH 95% CONFIDENCE INTERVAL.

Classification model		SER		Spk-ID
		Emo-DB	FAU-AEC	
		CA95%CI	CA95%CI	CA
MLGMM		66.81±0.44	60.00±0.13	76.50
Adapted GMM		79.48±0.31	61.09±0.12	83.08
SVM using	GMMIMK	85.62±0.29	62.48±0.07	88.54
	FK	87.05±0.24	61.54±0.11	88.54
	GMMSVK	87.18±0.29	59.78±0.19	87.93
	PSK	87.46±0.23	62.54±0.13	86.18
	CBSLPMK	87.60±0.20	61.04±0.09	84.85
	GUMIK	88.17±0.34	60.66±0.10	90.31
	GMMPMK	88.65±0.23	64.73±0.16	90.26
	GMMSLPMK	**92.24±0.19**	**67.96±0.10**	**91.35**

In our studies, the SVM-based classifiers using the CBSLPMK and GMMSLPMK are built using different values for Q corresponding the number of codewords and J corresponding to the number of levels in pyramid. In CBSLPMK, Q corresponds to number of clusters obtained using K-means clustering technique and in GMMSLPMK, Q corresponds to number of Gaussian components. The classification accuracies for the SVM-based classifier using CBSLPMK and GMMSLPMK are given in Table I for speech emotion recognition and speaker identification tasks. The best performances are shown using bold phase. It is seen that, the SVM-based classifiers using GMMSLPMK perform significantly better than SVM-based classifiers using CBSLPMK for all the tasks.

Table II compares the accuracies for speech emotion recognition and speaker identification tasks obtained using the GMM-based classifiers and SVM-based classifiers using the state-of-the-art dynamic kernels mentioned in Section II and the proposed SLPMKs. In this study, the GMMs whose parameters are estimated using the maximum likelihood (ML) method (MLGMM) and by adapting the parameters of the UBM or CIGMM to the data of a class (adapted GMM) [2] are considered to build GMM-based classifiers. The GMMs are built using the diagonal covariance matrices. The accuracies presented in Table II are the best accuracies observed among the GMM-based classifiers and SVM-based classifiers with dynamic kernels using different values for their parameters. The details of the experiments and the best values for the parameters can be found in [3] and [8]. It is seen that performance of the SVM-based classifiers using the state-of-the-art dynamic kernels is better than that of the GMM-based classifiers. It is also seen that the performance of the SVM-based classifiers using the proposed GMM-based SLPMK is significantly better than that of the SVM-based classifiers using the state-of-the-art dynamic kernels including the GMMPMK for speech emotion recognition and speaker identification tasks. The better performance of the SVM-based classifier using the proposed kernel is mainly due to the capabilities of the SLPMKs in capturing the local information better than the other dynamic kernels and also maintaining temporal information for some extent.

V. CONCLUSIONS

In this paper, we proposed the segment-level pyramid match kernels (SLPMKs) for the classification of varying length patterns represented as sets of feature vectors using the SVM-based classifiers. The SLPMK is computed by partitioning the speech signal into increasingly finer subparts and matching the corresponding subparts using a segment-level pyramid. The effectiveness of the proposed SLPMKs in building the SVM-based classifiers for classification of varying length patterns of long duration speech is demonstrated using studies on speech emotion recognition and speaker identification tasks. The performance of the SVM-based classifiers using the proposed GMM-based SLPMK is significantly better than the GMM-based classifiers and that of the SVM-based classifiers using the state-of-the-art dynamic kernels. The proposed SLPMK can be used for classification of varying length patterns extracted from video, audio, music, and so on, represented as sets of feature vectors using SVM-based classifiers. The proposed SLPMK is also useful for matching a pair of audio sequences in tasks such as audio retrieval, music retrieval etc.

REFERENCES

[1] Douglas A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, August 1995.

[2] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, January 2000.

[3] A. D. Dileep and C. Chandra Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1421–1432, Aug 2014.

[4] N. Smith, M. Gales, and M. Niranjan, "Data-dependent kernels in SVM classification of speech patterns," Tech. Rep. CUED/F-INFENG/TR.387, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, U.K., April 2001.

[5] K-A. Lee, C.H. You, H. Li, and T. Kinnunen, "A GMM-based probabilistic sequence kernel for speaker verification," in *Proceedings of INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 294–297.

[6] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, April 2006.

[7] C. H. You, K. A. Lee, and H. Li, "An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 49–52, January 2009.

[8] A. D. Dileep and C. Chandra Sekhar, "Speaker recognition using pyramid match kernel based support vector machines," *Internatiional Journel for Speech Technology.*, vol. 15, no. 3, pp. 365–379, September 2012.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, New York, NY, USA, June 2006, vol. 2, pp. 2169–2178.

[10] S. Boughorbel, J. P. Tarel, and N. Boujemaa, "The intermediate matching kernel for image local features," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2005)*, Montreal, Canada, July 2005, pp. 889–894.

[11] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *The Journal of Machine Learning Research*, vol. 8, pp. 725–760, 2007.

[12] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, April 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, and W. S. B. Weiss, "A database of German emotional speech," in *Proceedings of INTERSPEECH*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[14] S. Steidl, "Automatic classification of emotion-related user states in spontaneous childern's speech," PhD thesis, Der Technischen Fakultät der Universität Erlangen-Nürnberg, Germany, 2009.

[15] "The NIST year 2002 speaker recognition evaluation plan," *http://www.itl.nist.gov/iad/mig/tests/spk/2002/*, 2002.

[16] "The NIST year 2003 speaker recognition evaluation plan," *http://www.itl.nist.gov/iad/mig/tests/sre/2003/*, 2003.