Interspeech 2018
2-6 September 2018, Hyderabad

Cross-lingual Speech Emotion Recognition through Factor Analysis

Brecht Desplanques, Kris Demuynck

Ghent University - imec, IDLab, Department of Electronics and Information Systems, Belgium
Brecht.Desplanques@UGent.be, Kris.Demuynck@UGent.be

Abstract

Conventional speech emotion recognition based on the extraction of high level descriptors emerging from low level descriptors seldom delivers promising results in cross-corpus experiments. Therefore it might not perform well in real-life applications. Factor analysis, proven in the fields of language identification and speaker verification, could clear a path towards more robust emotion recognition. This paper proposes an iVector-based approach operating on acoustic MFCC features with a separate modeling of the speaker and emotion variabilities respectively. The speech analysis extracts two fixed-length low-dimensional feature vectors corresponding to the two mentioned sources of variation. To model the speaker-related nuisance variability speaker factors are extracted using an eigenvoice matrix. After compensating for this speaker variability in the supervector space, the emotion factors (one per targeted emotion) are extracted using an emotion variability matrix. The emotion factors are then fed to a basic emotion classifier. Leave-one-speaker-out cross-validation on the Berlin Database of Emotional Speech EMO-DB (German) and IEMOCAP (English) datasets lead to results that are competitive with the current state-of-the-art. Cross-lingual experiments demonstrate the excellent robustness of the method: the classification accuracies degrade less than 15% relative when emotion models are trained on one corpus and tested on the other.

Index Terms: emotion recognition, factor analysis, emotion factor extraction, cross-lingual

1. Introduction

Automatic recognition of paralinguistic information in speech has gained significant attention over the last couple of years. This is illustrated by the yearly Computational Paralinguistics ChallengEs (ComParE) [1, 2]. In this paper we tackle the automatic emotion recognition in the context of telehomecare. A Flemish call center regularly contacts people who are socially isolated or in need of nursing care. The goal is to make an assessment of their health and mental state, and to timely intervene when necessary. An acoustic voice analysis of the telephone conversations can deliver crucial information about the individual's well-being, feelings of loneliness, mental state, pain problems, etc.

Supervised learning requires training data with paralinguistic labels, and such data is rarely available in Flemish. Therefore we rely on language-independent factor analysis techniques [3, 4] popularized in automatic language recognition and speaker verification to build an emotion recognition system. The robustness of the method is verified with cross-lingual experiments. Compared to the predominant approach [2] with high-dimensional feature vectors computed using statistical functionals [5], the proposed factor analysis approach extracts a very low-dimensional emotion factor vector. This paves the way for online and speaker-adaptive approaches [6] in our telehomecare use case.

2. System setup

A favored approach in automatic language recognition to extract information from a speech utterance is that of iVectors [4] or Total Variability (TV) modeling. All variability is modeled in a single low dimensional subspace. In the emotion recognition domain a low rank rectangular matrix $\boldsymbol{T}$, called the TV matrix or the iVector extractor, can be used to approximate the GMM mean supervector $\boldsymbol{m}_{i,s}$ of a segment i uttered by speaker s as

$$\boldsymbol{m}_{i,s} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{x}_{i,s} \tag{1}$$

with $\boldsymbol{m}$ being the Universal Background Model (UBM) supervector and $\boldsymbol{x}_{i,s}$ being the fixed-length iVector that contains all the information concerning speaker s and the affective speaker state (emotion) of that speaker in segment i. The UBM is trained on frame-based acoustic features. In the final stage a classifier, e. g. a Gaussian Backend [7] extracts the speaker state information from the iVector.

However, in order to build a robust iVector extractor and emotion classifier, training data containing a sufficient number of different speakers acting in a wide range of affective speaker states is needed. This poses a serious challenge. Moreover, utterances produced by the same speaker in a varying emotional state should in principle deliver identical speaker information and iVectors across these utterances share redundant speaker information. This leads us to an approach in which the speaker and the affective speaker state variability are treated separately [8]. This separation of variability sources has been shown to deliver state-of-the-art performance in the domain of language identification [6]. In this paper a speaker variability eigenvoice model is trained on a large out-of-domain dataset which only needs speaker label annotations. The eigenvoice model is then used to train a speaker-independent emotion variability model on the specialized emotion dataset.

2.1. Emotion factor extraction

During evaluation we pool together all data of a particular speaker s and use an eigenvoice matrix $\boldsymbol{U}$ to extract the relevant speaker factors $\boldsymbol{y}_s$ in a way similar to the iVector framework [9]:

$$\boldsymbol{m}_s = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{y}_s \tag{2}$$

This operation shifts the UBM towards a speaker-dependent GMM according to the triggered eigenvoices in $\boldsymbol{U}$. Next, we extract emotion factors $\boldsymbol{y}_i$ for each test segment i uttered by speaker $s(i)$:

$$\boldsymbol{m}_i = \boldsymbol{m}_{s(i)} + \boldsymbol{V}\boldsymbol{y}_i \tag{3}$$

The corresponding speaker-dependent GMM is shifted towards a speaker-dependent and emotion-dependent GMM matching the test utterance within the emotion subspace defined by emotion variability matrix $\boldsymbol{V}$. All affective speaker state information is conveyed by the emotion factors $\boldsymbol{y}_i$, which can be extracted by a simple classifier due to the low dimensionality of this vector.

10.21437/Interspeech.2018-1778

Note that we could have used Joint Factor Analysis (JFA) [3] to extract the speaker factors and emotion factors simultaneously for all utterances of a particular speaker. However, this simultaneous extraction is computationally much more demanding and did not return better results. Moreover, in an online use case it might be useful to fix the speaker factors on a previous set of recordings of that speaker, which is straightforward to implement in our two-step procedure. In the subsequent subsections we will discuss how to build all the necessary subspace models and emotion classifiers.

2.2. Speaker variability modeling

As no emotion annotations are needed, the UBM and eigenvoice matrix can be trained on a large out-of-domain dataset with a large collection of different speakers. The eigenvoice matrix $\boldsymbol{U}$ is constructed by means of Principal Component Analysis (PCA) initialization [10] followed by iterating the non-simplified Expectation-Maximization (EM) algorithm described in [11] until it converges. However, we want the speaker factors to react to speaker changes only and not to intra-speaker variability due to changes in the channel or the background. Thus, during the training of the eigenvoice matrix $\boldsymbol{U}$ we pool together all turns of a certain speaker into one instance of that speaker, meaning that the channel and background variability are incorporated in the speaker model.

2.3. Emotion variability modeling

In this section we discuss the training procedure of emotion variability matrix $\boldsymbol{V}$ in greater detail.

2.3.1. Speaker-compensated Baum-Welch statistics

The mathematical procedure of extracting latent factors relies on the estimation of the zero- and centralized first order Baum-Welch statistics [11] estimated with the UBM. In previous work [6] we have shown that by averaging the centralized first order statistics across speakers within a language class one can build a robust language variability model. Due to the small number of speakers in our emotion datasets this strategy will not be sufficient to eliminate all nuisance speaker variability. We propose to centralize the first order statistics around a speaker-dependent supervector, instead of the UBM supervector. This suppresses the speaker dependencies per segment before eliminating the remaining variability via the standard approach of averaging the statistics across utterances within the considered speaker state.

These statistics of a given utterance i made by speaker $s(i)$ corresponding with time interval $\mathcal{T}(i)$ are estimated as:

$$N_i^{(m)} = \sum_{\forall t \in \mathcal{T}(i)} \gamma_t^{(m)} \tag{4}$$

$$\boldsymbol{f}_i^{(m)} = \sum_{\forall t \in \mathcal{T}(i)} \gamma_t^{(m)} \boldsymbol{o}_t - N_i^{(m)} \boldsymbol{m}_{s(i)}^{(m)} \tag{5}$$

$$\boldsymbol{m}_s^{(m)} = \boldsymbol{m}^{(m)} + \boldsymbol{U}^{(m)} \boldsymbol{x}_s \tag{6}$$

In these equations, $\boldsymbol{o}_t$ is the feature vector at time t and $\gamma_t^{(m)}$ is the occupation probability of mixture m according to the UBM at that time. $\boldsymbol{m}^{(m)}$ and $\boldsymbol{U}^{(m)}$ are the components corresponding with mixture m of the UBM supervector and eigenvoice matrix respectively . The first order statistics defined in (5) are centralized around the speaker-dependent supervector $\boldsymbol{m}_s$ instead of the UBM supervector. Note that $\boldsymbol{m}_s$ is estimated on all data belonging to speaker s.

2.3.2. Training the emotion variability matrix

Since the number of emotions is small, there is no need to rely on the EM algorithm for finding a compact representation $\boldsymbol{V}$ of the emotion subspace. Instead, we assign one vector directly to each of the emotions and set the values of the column vectors $\boldsymbol{V}_e$ of $\boldsymbol{V}$ equal to the offset between the ML supervector $\boldsymbol{m}_e$ of the corresponding emotion e and the corresponding speaker-adapted supervectors $\boldsymbol{m}_s$. This is achieved by averaging the speaker-compensated first order Baum-Welch statistics across utterances within a speaker state per mixture m:

$$\boldsymbol{V}_e^{(m)} = \frac{\sum_{\forall i \in e} \boldsymbol{f}_i^{(m)}}{\sum_{\forall i \in e} N_i^{(m)}} \tag{7}$$

This ensures that the speaker-adapted GMMs can be shifted towards speaker and emotion dependent GMMs when performing adaptation with matrix $\boldsymbol{V}$.

2.3.3. Removal of the neutral emotion shift

The evaluation datasets treat neutral speech as an one of the emotion categories. However, our UBM should model neutral speech already, which calls the validity of column $\boldsymbol{V}_n$ in $\boldsymbol{V}$ corresponding with a shift towards neutral speech into question. Due to the small number of training speakers, this shift is expected to be mostly dataset-specific and it might behave unexpectedly on unseen conditions. We propose to eliminate this redundant shift in a second training iteration. First the UBM supervector is adapted towards the neutral speech:

$$\boldsymbol{m}^* = \boldsymbol{m} + \boldsymbol{V}_n \tag{8}$$

Next, the speaker factors $\boldsymbol{x}_s$ are re-extracted and the speaker-compensated Baum-Welch statistics are re-estimated given the occupation probabilities emitted by the adapted UBM. The emotion variability matrix $\boldsymbol{V}$ is reconstructed according to the training procedure explained in Section 2.3.2. Finally, the remaining neutral emotion shift $\boldsymbol{V}_n$ is removed from $\boldsymbol{V}$ and the other emotions shifts are updated accordingly:

$$\boldsymbol{V}_e^* = \boldsymbol{V}_e - \boldsymbol{V}_n \tag{9}$$

This final procedure defines the origin of the emotion shifts $\boldsymbol{V}_e^*$ in $\boldsymbol{V}^*$ as the supervector of the neutral speech in the emotion dataset. Given the updated models, new emotion factors $\boldsymbol{y}_i^*$ are extracted per utterance.

2.4. Emotion factor classification

A simple Gaussian Backend (GB) classifier models the distribution of the emotion factors $\boldsymbol{y}^*$ of the target emotion e by means of a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_e, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ a full covariance matrix shared by all target emotions [7]. The classification is based on the following emotion score:

$$a_{e,i}^* = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_e)^T \boldsymbol{y}_i^* - \frac{1}{2}\boldsymbol{\mu}_e^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_e \tag{10}$$

The emotion producing the highest score $a_{e,i}^*$ is selected.

2.5. Acoustic features

The acoustic inputs of the UBM are shifted delta cepstral (SDC) MFCC feature vectors [12] on 10ms frames. They are acknowledged to constitute a richer representation of the signal dynamics than the standard dynamical features derived from the

MFCCs and have been shown to significantly improve performance in the domain of emotion recognition [13]. SDC features are defined by four parameters: N, d, P and k. The first $(2k+1)N$ features of frame t consist of the Δs of the N static MFCCs $c_1 \dots c_N$, computed for frames $t + iP$ ($i = -k \dots k$). A Δ at frame t is computed over the window $(t-d, t+d)$. We use a standard configuration of $N = 10$, $d = 2$, $P = 3$ and $k = 2$. Supplementing the SDCs with 16 static MFCCs (the $c_1 \dots c_{16}$ of frame t) and a normalized log-energy finally leads to a feature vector of dimension 67. The normalized log-energy component is defined as:

$$\log E_{\text{nrm}}(t) = \log E(t) - \overline{\log E(t)} \tag{11}$$

It is equal to zero when the log-energy is equal to a running mean log-energy $\overline{\log E(t)}$ and positive when it is larger. The running mean is computed by means of a leaky integrator with a time constant of 5 seconds.

We do not incorporate any kind of frame-based voice activity detection (VAD) or additional speaker-based feature normalization (e. g. cepstral mean subtraction). Note that the SDC features span a relatively large window and features of low-energy non-speech frames can contain information about surrounding speech frames. Utterance-based MFCC feature normalization to compensate for channel effects significantly degraded the emotion classification performance and was not incorporated as well. Future work will focus on finding ways to eliminate the channel variability without removing too much emotion-specific information and integrating extra pitch information in the proposed approach.

2.6. Baseline system

The performance of our proposed factor analysis technique will be compared with the conventional approach of extracting high-dimensional feature vectors using statistical functionals on low-level descriptors [2]. We use the openSMILE [5] toolbox and the *emobase.conf* configuration file to extract 988 features. Heuristic feature selection is applied to retain the 70 features that deliver optimal classification performance on the evaluation sets. All baseline experiments use the same selection of features. We apply speaker normalization on the functional level so that the feature vectors have a mean of zero and standard deviation of one for each speaker. An ensemble of 5 Extreme Learning Machines (ELM) [14] with each 200 hidden nodes is used to classify the feature vectors into emotion categories. The ELM is defined as a feed-forward neural network (a Multi-Layer Perceptron) with a randomly fixed hidden layer and a linear output layer whose weights are fixed to minimize the cross entropy between the computed and the desired outputs. It is mathematically proven that the ELM is as powerful as a fully trained Multi-Layer Perceptron [14].

3. Experimental results

3.1. Datasets

3.1.1. Emotion datasets

The German Berlin Database of Emotional Speech EMO-DB [15] includes 535 short utterances from seven basic emotions (anger, boredom, disgust, fear, happiness, sadness, neutral). In an anechoic chamber, ten native German professional actors expressed ten different sentences. Each sentence was expressed in all emotions. Sentences leading to annotator disagreement were omitted.

We also conducted experiments on the English Interactive Emotion Dyadic Motion Capture (IEMOCAP) dataset [16]. This dataset contains approximately 12 hours of audio-visual data from five mixed-gender pairs of actors, in this paper we only focus on the audio data. Each interactive session lasts about 5 minutes and is based on either a scripted or improvised scenario. The sessions were manually segmented into utterances. Each utterance was annotated by at least 3 annotators into categorical labels. We examine the anger, happiness, excitation, neutrality and sadness emotion classes for the 5531 utterances that had majority consensus across the annotators. The classes of happiness and excitation are merged into a single class. The class distribution is: 20.0% angry, 19.6% sad, 29.5% happy, and 30.9% neutral.

3.1.2. Out-of-domain datasets

The English 1996 HUB4 Broadcast News [17] data (66 hours, 3009 speakers) is used to train an English UBM and eigenvoice model. We harvested 15 hours of speech from ZDF podcasts[1] (1106 speakers) as German training data for these models. Speaker labels for the podcasts were auto-generated by our speaker diarization system [18].

3.2. Evaluation protocol

The emotion recognition experiments are based on 10-fold leave-one-speaker-out cross-validation (CV). First, an UBM and eigenvoice model are trained on the out-of-domain dataset that matches the language of the evaluation dataset. Next, the emotion variability matrix estimation is performed on the 9 training speakers. The system is evaluated on the remaining speaker. This process is repeated for each speaker in the dataset.

In both evaluation sets the emotion class distribution is not too heavily skewed and we rely on the (weighted) accuracy (i. e. the probability that a test utterance is classified correctly) to evaluate the different system setups.

3.3. Intra-corpus experiments

The number of UBM mixtures is set to 64. The rank of the eigenvoice matrix $\boldsymbol{U}$ is 50. The results compared to the baseline OpenSMILE approach can be found in Table 1. We included both GB and ELM classification of the extracted emotion factors. We reduced the number of ELM hidden nodes to 20 to match the low dimensionality of this input.

Table 1: *Emotion recognition performance in accuracy(%).*

dataset	EMO-DB	IEMOCAP
OpenSMILE + ELM (200 nodes)	83.2	54.8
Emotion Factors + GB	82.8	55.2
Emotion Factors + ELM (20 nodes)	81.3	55.4
OpenSMILE Emotion Factors + ELM (200 nodes)	85.6	56.1

Both feature extraction approaches achieve similar performance and the results reported in Table 1 are competitive with the results reported in literature [19, 8, 20, 21]. The emotion factor extraction is clearly successful in reducing the feature dimensionality towards a feature space that is directly interpretable, which makes the simple GB a suitable classifier. It also suggests that the adaptation techniques proposed in [6]

[1] http://www.zdf.de

could lead to an adaptive strategy for emotion recognition. ELM classification on the speaker-normalized concatenation of the OpenSMILE features and the extracted emotion factors delivers better performance than the standalone systems, which indicates the two approaches contain complementary information. These results of feature fusion are shown in the final row of Table 1.

3.4. Cross-corpus experiments

To evaluate the robustness of our proposed emotion recognition system and to assess real-word performance, we train the emotion models on one corpus and evaluate the system on the other. Another argument for this cross-lingual experiment is the fact that annotated emotional speech is hard to collect, and one may need to resort to non-target languages to collect sufficient data. English, Flemish and German are Germanic languages and share a very similar cultural background, we therefore assume that the emotion definitions are transferable across the datasets. We select the 4 matching emotions in the IEMOCAP and EMO-DB datasets (*angry*, *happy*, *neutral* and *sad*).

For the proposed emotion factor extraction, we train the UBM and eigenvoice model $\boldsymbol{U}$ on the out-of-domain dataset that has the same language as the target evaluation set. Such data should be easily obtainable as there is no need for emotion labels. Finally, the emotion variability model is trained on the non-target evaluation set. In case of the baseline approach, we reduce the number of hidden nodes to 20 in order to enhance the generalization performance across datasets. The ELM is simply trained on the non-target dataset and evaluated on the target one. The results are shown in Table 2. The corresponding intra-corpus CV experiments with identical system settings are also included. Note the improved intra-corpus performance on EMO-DB due to the reduced number of emotion classes.

Table 2: *Intra-corpus and cross-lingual emotion recognition performance in accuracy(%) on 4 emotion classes (happy, angry, neutral and sad).*

dataset	EMO-DB (4 emotions)	IEMOCAP
	intra-corpus	
OpenSMILE + ELM (200 nodes)	90.3	54.8
OpenSMILE + ELM (20 nodes)	88.5	51.3
Emotion Factors + GB	90.5	55.2
	cross-corpus	
OpenSMILE + ELM (200 nodes)	51.9	38.9
OpenSMILE + ELM (20 nodes)	61.0	40.9
Emotion Factors + GB	81.4	48.4

The OpenSMILE system underperforms significantly in the cross-corpus experiments and it experiences performance drops of up to 40% in absolute terms. The reduction of the number of hidden nodes in the ELM increases its generalization performance slightly at the cost of intra-corpus performance. The relative cross-corpus performance degradation of the proposed emotion factor extraction on the other hand, is about 12.5% only. This is significantly less than the impact reported in [22, 21]. Extra experiments on feature fusion did not result in enhanced results, probably due to the degraded performance of the baseline system.

3.5. Extra analysis IEMOCAP results

The performance on the IEMOCAP data significantly lags the performance on the EMO-DB data, which may be partially explained by the fact that IEMOCAP contains spontaneous, less over-acted recordings. In order to make a more useful analysis of the IEMOCAP performance we suggest to perform an oracle segmentation and group all consecutive utterances (disregarding utterances of out-of-set emotions) of identical emotion for each speaker. Each group is classified as one unit. This is a viable approach as the utterances are part of 5 minute scenarios with one dominant emotion. The evaluation of the emotion factor extraction remains utterance-based. The accuracy increases from 55.2% to 70.7% when oracle segmentation is enabled. This indicates that systems exploiting contextual information by looking for speaker state changes in an initial stage could drastically outperform the current utterance-based classification approaches. We note that the average duration per test unit increased from 4.5s to 23s. A crude approach where we include the two previous and the two subsequent utterances of the same speaker during the emotion factor extraction of the considered test utterance already results in an accuracy of 61.5% on the IEMOCAP data. To allow for a varying dominant emotion in the conversations, we did not take conversations (file) boundaries into account and concatenated all utterances of the test speaker.

We also note that there is a big discrepancy between the performance on the scripted and improvised subset of the IEMOCAP data. The results of training and evaluating the emotion factor extraction on each of the subsets with 10-fold CV can be found in Table 3. A possible explanation for this performance gap might lay in the fact that it comes more natural to an actor to convey emotions when he/she is not restricted by a script.

Table 3: *Emotion recognition performance in accuracy(%) on different subsets of IEMOCAP.*

IEMOCAP subset	improvised	scripted
utterance-based	61.4	51.0
oracle segmentation	73.1	65.9

4. Conclusion

In this work we proposed a two-step factor analysis approach for speech emotion recognition. After a reduction of the speaker variability in utterance-based supervectors by an eigenvoice analysis, the remaining variability is projected to an affective speaker state subspace. The coordinates in this low-dimensional space can be directly interpreted as a score per emotion, which paves the way for speaker-adaptive applications. The approach delivers competitive results for matching training and test conditions compared to the conventional emotion classification based on high-dimensional feature vectors computed using statistical functionals on lower level descriptors. Cross-lingual experiments on English and German data illustrate the improved robustness of the proposed approach which bodes well for real-world applications of speech emotion recognition.

5. Acknowledgements

We would like to thank Prof. Shrikanth S. Narayanan for sharing the IEMOCAP data. This work was supported by KMO Innovatie VLAIO project No. 150339 - Global Telehomecare.

6. References

[1] B. W. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hnig, J. R. Orozco-Arroyave, E. Nth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition." in *Proc. Interspeech*, 2015, pp. 478–482.

[2] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. Interspeech*, 2016, pp. 2001–2005.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. 21st ACM International Conference on Multimedia*, 2013, pp. 835–838.

[6] B. Desplanques, K. Demuynck, and J.-P. Martens, "Robust language recognition via adaptive language factor extraction," in *Proc. Interspeech*, 2014, pp. 2160–2164.

[7] D. Martìnez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Proc. Interspeech*, 2011, pp. 861–864.

[8] M. Li, A. Metallinou, D. Bone, and S. S. Narayanan, "Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling," in *Proc. ICASSP*, 2012, pp. 1937–1940.

[9] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[10] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J. Černocký, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

[11] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proc. ICASSP*, 2011, pp. 4516–4519.

[12] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, and J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. ICSLP*, 2002, pp. 89–92.

[13] G. Liu, Y. Lei, and J. H. Hansen, "A novel feature extraction strategy for multi-stream robust emotion identification," in *Proc. Interspeech*, 2010, pp. 482–485.

[14] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, 2006.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[17] J. Garofolo, J. Fiscus, and W. Fisher, "Design and preparation of the 1996 hub-4 broadcast news benchmark test corpora," in *Proc. of DARPA Speech Recognition Workshop*, 1997, pp. 15–21.

[18] B. Desplanques, K. Demuynck, and J. Martens, "Adaptive speaker diarization of broadcast news based on factor analysis," *Computer Speech & Language*, vol. 46, pp. 72–93, 2017.

[19] T. Chaspari, D. Dimitriadis, and P. Maragos, "Emotion classification of speech using modulation features," in *Proc. EUSIPCO*, 2014, pp. 1552–1556.

[20] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, 2014, pp. 223–227.

[21] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," in *Proc. Interspeech*, 2017, pp. 1113–1117.

[22] J. H. Jeon, D. Le, R. Xia, and Y. Liu, "A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception," in *Proc. Interspeech*, 2013, pp. 2837–2840.