Predicting Speaker Recognition Reliability by Considering Emotional Content

Srinivas Parthasarathy
The University of Texas at Dallas
Email: sxp120931@utdallas.edu

Carlos Busso
The University of Texas at Dallas
Email: busso@utdallas.edu

***Abstract*—Studies have shown that emotional variability in speech degrades the performance of speaker recognition tasks. Of particular interest is the error produced due to mismatch between training speaker recognition models with neutral speech and testing them with expressive speech. While previous studies have considered categorical emotions, expressive speech during human interaction conveys subtle behaviors that are better characterized with continuous descriptors (e.g., attributes such as arousal, valence, dominance). As the emotion becomes more intense, we expect the performance of speaker recognition tasks to drop. Can we define emotional regions for which the speaker recognition performance is expected to be reliable? This study focuses on automatically predicting reliable regions for speaker recognition by analyzing and predicting the emotional content. We collected a unique emotional database from 80 speakers. We estimate speaker recognition performance as a function of arousal and valence, creating regions in this space where we can reliably recognize the identity of a speaker. Then, we train speech emotion recognizers designed to predict whether the emotional content in a sentence is within the reliable region. The experimental evaluation demonstrates that sentences that are classified as reliable for speaker recognition tasks have lower *equal error rate* (EER) than sentences that are considered unreliable.**

1. Introduction

An important area in speech processing is speaker recognition, where the task is to determine the identity of a speaker from a pool of individuals [1]. The process consists of training speaker models, which are evaluated with separate recordings (test partition). There are several conditions that lead to a drop in speaker recognition performance including emotions. Speaker recognition models are trained with fairly emotionally-neutral speech. If the test speech is emotional, the speech features will deviate from their expected values, creating a mismatch between train and test conditions. This mismatch affects the performance of speaker recognition systems [2], [3], [4], [5], [6], [7].

Previous studies have shown the effect of emotion in speaker recognition tasks, providing compensation schemes [8], [9], [10], [11]. However, these compensation approaches may negatively affect the performance of speaker recognition systems for emotionally neutral recordings [3]. It is important to identify when emotion compensation scheme are needed. In some cases, the speaker recognition performance will be so unreliable that it is better to discard that recording, prioritizing forensic analysis on other recordings. Therefore, it is important to distinguish the emotional range conveyed on speech for which we can reliably recognize the identity of the speaker.

This study (1) provides a comprehensive analysis of the reliability of a speaker recognition task in the presence of expressive speech, and (2) uses emotional classifiers to determine sentences for which the speaker recognition performance is less reliable. First, the study analyzes the performance of a speaker recognition system evaluated with expressive speech. The analysis is conducted on a subset from the MSP-PODCAST corpus [12]. The corpus contains several hours of natural emotional speech from multiple speakers appearing on audio-sharing websites. The emotional content of the corpus is annotated in terms of the following emotional attributes: arousal (very calm versus very active), valence (very negative versus very positive) and dominance (very weak versus very strong). We analyze the speaker recognition performance as a function of the emotional attribute scores, extending our previous work from 40 to 80 speakers [13]. From the analysis, we identify regions/boundaries in the arousal-valence space, defining three classes for reliable, uncertain and unreliable sentences (i.e., we can/cannot recognize the identity of a speaker within a certain threshold of error). We formulate this problem using a multiclass emotional classifier, where the classes are the regions of reliability in the emotional space. The challenge is to automatically detect when utterances fall in these regions.

We evaluate performance in terms of *equal error rate* (EER) of all utterances in the predicted region (reliable, uncertain and unreliable). Our results show that the EERs in the predicted classes match the EERs obtained when we use the emotional labels assigned to the sentences. The results validate our framework for predicting reliability of the speaker recognition task in the presence of expressive speech. The key contribution of this study is combining emotion recognition and speaker recognition using a novel framework, providing a valuable tool for forensic analysis.

2. Related Work

2.1. Speaker Recognition and Emotion

Emotional speech affects the performance of speaker recognition systems [2], [3], [4], [5], [6], [7]. Previ-

978-1-5386-0563-9/17/$31.00

ous studies have commonly analyzed speaker verification/recognition in terms of categorical emotions such as happiness, anger and sadness [3], [6], where the goal is to create compensation techniques to improve speaker recognition performance. These methods include modification of features from neutral to emotional categories [4], [6], and use of emotional and gender information to train the system [10]. These studies have two main drawbacks. First, they have represented emotion with categorical emotions, which is not very practical, as emotions in daily interaction includes ambiguous behaviors with mixed emotions [14]. Categorical labels do not capture the intensity or within-class variability which may have an important effect on speaker recognition tasks (e.g., *hot anger* may affect speaker verification performance while *cool anger* may have no effect). Second, previous studies have used a limited number of speakers, who were asked to provide acted recordings. Our study attempts to break these barriers.

In our previous work, we evaluated the speaker verification performance in terms of emotional attributes using a set of 40 speakers from the MSP-PODCAST corpus [13]. Instead of emotional categories, the evaluation represented emotions using arousal, valence and dominance. This representation provides better resolution to study the role of emotion in speaker recognition tasks (note that emotional categories can be mapped to attribute dimensions [15]). The analysis showed that the EER dropped as the emotional attributes depart from neutral speech. This paper follows up with the analysis, increasing the number of speakers from 40 to 80 (Sec. 3.3). The analysis is used to train an emotion classifier to predict the reliability of speaker recognition tasks.

2.2. Predicting Speaker Recognition Reliability

An important problem in speaker recognition is to predict the reliability of the results under various conditions. Previous studies have used a confidence measure to improve the speaker recognition model. Huggins and Grieco [16] used a confidence measure which included several factors that affect speaker identification tasks (mismatch between train and test data quality in terms of *signal to noise ratio* (SNR), duration, number of speakers). The addition of the model confidence measure reduced speaker recognition errors by 2.8%. Campbell et al. [17] measured the confidence of speaker verification for forensic tasks, using a regression model trained with meta data (e.g., utterance duration, channel information and SNR). Richiardi et al. [18] used a probabilistic measure to evaluate the reliability of speaker verification under noisy conditions. They proposed a Bayesian network that takes as input the speaker verification likelihoods and the SNR, predicting the reliability of a given sentence. By discarding unreliable sentences, the system reduced its EER from 9.3% to 2.8%. Villalba et al. [19] analyzed various speech quality measures to predict the reliability of a speaker verification task, showing that the best features were the modulation index, SNR and *vector Taylor series* (VTS) coefficients to linearly approximate the non-linear effect of noise and reverberation on the *Mel-frequency cepstral coefficients* (MFCCs). To the best of our knowledge, this is the first study predicting the reliability of speaker recognition task in terms of emotional speech.

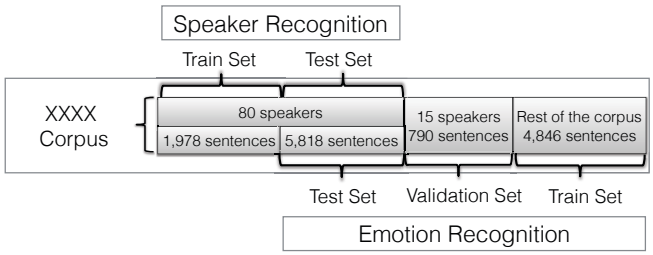


Figure 1. Data partition for speaker and emotion recognition tasks. This partition (1) unifies the test set for both tasks, and (2) gives our best effort to establish speaker independent sets for emotion recognition.

3. Speaker Recognition for Expressive Speech

3.1. Database

This study uses the MSP-PODCAST corpus [12]. This corpus is an extensive collection of spontaneous speech from multiple speakers appearing in Creative Commons licensed recordings downloaded from audio-sharing websites. Some key aspects of the corpus are the different conditions in which the recordings are collected, large number of speakers, and natural content from spontaneous conversations conveying emotional behaviors. Each utterance in the corpus is annotated by at least five raters using an online crowdsourcing platform using a protocol inspired in Burmania et al. [20]. Emotional dimensions are annotated using seven-Likert scales for arousal (1- very passive versus 7- very active), valence (1- very negative versus 7- very positive) and dominance (1- very weak versus 7- very strong). The sentences are also annotated for primary categorical emotions where raters selected the class that best represented the utterance. The corpus currently contains 13,432 sentences (21h 15m). The readers are referred to Lotfian and Busso [12] for more information about the corpus.

For the speaker recognition task, we manually annotate the identity of the speakers in the database. We have identified 95 speakers from whom we have at least 300s of speech data. Figure 1 illustrates the data partition of the database for speaker and emotion recognition tasks. The speaker recognition task is evaluated with sentences from 80 speakers (7,796 – 13h 21m), leaving enough samples outside this set to train emotion classifiers (790 for validation set, 4,846 for train set). The partition gives our best effort to have speaker-independent partitions for emotion recognition experiments. Notice that the test set for both tasks have to be the same to assess the effectiveness of emotion recognition to assess the reliability in speaker recognition tasks.

3.2. Speaker Recognition Framework

The speaker recognition system uses the i-vector framework with a mean normalized *probabilistic linear discrim-*

TABLE 1. CRITERIA TO FORM THE TRAIN SET FOR THE SPEAKER RECOGNITION TASK, USING THE MSP-PODCAST DATABASE.

Criteria
CRITERION 1: Add utterances at random where the categorical emotion is "neutral" and arousal, valence, and dominance values are inside the range [3,5]. **1203 utterances**
CRITERION 2: Add utterances at random where the categorical emotion is "neutral" and arousal, valence, and dominance values are inside the range [2,6]. **417 utterances**
CRITERION 3: Add utterances at random where the categorical emotion is "neutral" and arousal, valence, and dominance values are inside the range [1,7]. **7 utterances**
CRITERION 4: Add utterances at random where the arousal, valence, and dominance values are inside the range [3,5], regardless of the categorical emotion. **206 utterances**
CRITERION 5: Add utterances at random where the arousal, valence, and dominance values are inside the range [2,6], regardless of the categorical emotion. **145 utterances**
[The range for attributes is [1-7], where 4 is neutral value.]

inant analysis (PLDA) back-end. The i-vector model provides a method for compressing high dimensional Gaussian super-vectors into a low dimension space [21], [22], [23]. In the PLDA framework, the average of all enrollment i-vectors is used as the final representation of the speaker model [24]. More details about the framework can be found in Parthasarathy et al. [13]. We use a 256 component mixture UBM for training the speaker recognition models. The models are trained on a 39 dimensional feature vector consisting of 13 MFCCs + Δ + $\Delta\Delta$. The dimension of the i-vector is empirically set to 200.

To understand the effect of emotional speech on the speaker recognition task, we create a mismatch where the models are trained with neutral speech, and tested with either neutral or emotional speech. We have at least 300s of speech from 80 speakers in the corpus, where we used 150s for training the models. Since the distribution of emotion varies across speakers, defining 150s of *neutral* speech is not straightforward. Table 1 shows the criteria used to sequentially get 150s of speech per speaker to train our models. Table 1 also gives the total number of utterances under each criterion. We use 1,978 sentences to train the models. The remainder 5,818 sentences from the 80 speakers are used to test the models (Fig. 1).

3.3. Speaker Recognition results

We use EER to evaluate the speaker recognition performance. The analysis follows the approach presented by Parthasarathy et al. [13], which aims to understand speaker recognition performance as a function of arousal and valence values. We split the arousal-valence space into 2D bins separated by 0.1 (arousal and valence scores are in the range [1,7]). Each bin is associated with the sentences that are within a 0.4×0.4 window centered at the middle of the 2D bin. Unlike our previous study where the EER was individually calculated for each test utterance, we estimate a single EER value for the sentences associated with the bin. This EER value is assigned to the bin. To make the analysis more robust, we only consider bins with at least

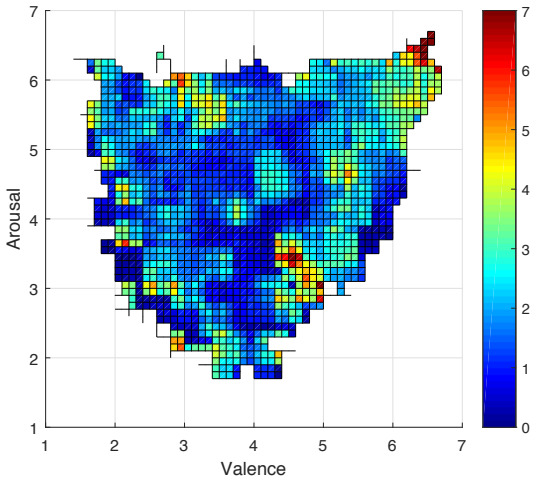


Figure 2. EER of speaker recognition task as a function of arousal and valence.

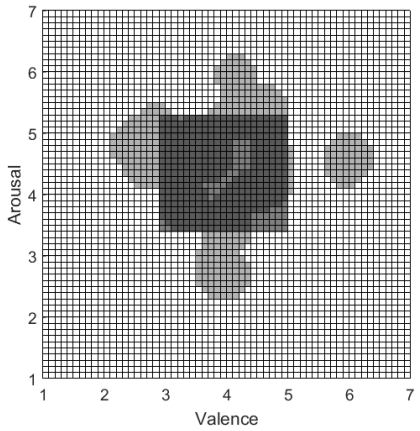
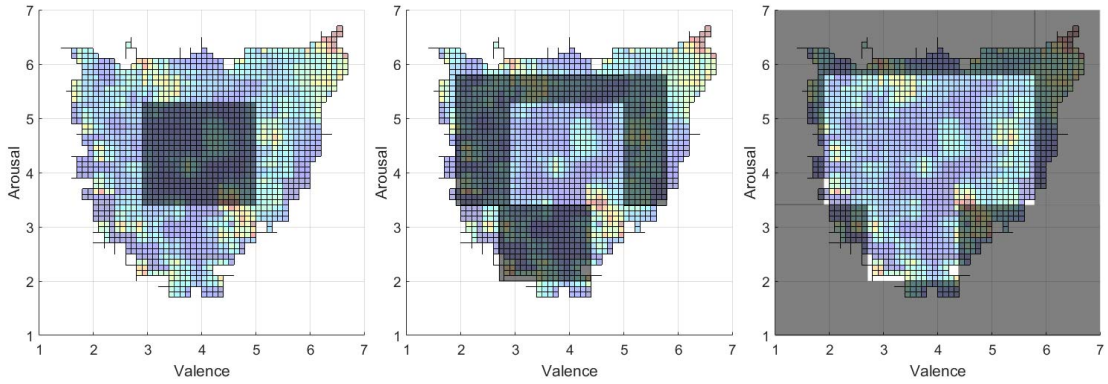


Figure 3. Binary map indicating the regions where we achieve less than 2% EER for the speaker recognition task after dilation and erosion. Bins in black indicate EER less than 2%.

10 speakers with at least 1 test utterance, increasing the diversity across speakers and sentences. Figure 2 illustrates the mapping between the emotional attributes and the EER for the speaker recognition task. White spaces indicate that a given region does not meet our criteria. Figure 2 shows lower EER around the neutral region corresponding to the [4,4] coordinate. The EER increases as we deviate from the neutral region. The results are similar to our previous study, but more conclusive as our current analysis is with twice as many speakers (i.e., 80 speakers). There are some regions with lower EER around the boundaries, which is probably an artifact of having lower number of utterances and speakers around these regions. Importantly, the figure identifies regions of reliability for the speaker recognition task (areas in the arousal-valence space where the EER is small). The rest of the study focuses on predicting the speaker recognition reliability using emotional speech classifiers.



(a) Relieable Region (b) Uncertain Region (c) Unrelieable Region

Figure 4. Region of reliability for speaker recognition task. Shaded gray regions indicate the boundaries for reliable, uncertain and unreliable classes chosen for the evaluation in Section 4.1.

4. Reliability for Speaker Recognition

This section formulates the prediction of the reliability in speaker recognition tasks as a speech emotion classification problem. Section 4.1 explains the framework to define regions of reliability for speaker recognition tasks. Section 4.2 introduces the emotional classifiers trained to automatically predict these regions of reliability.

4.1. Regions of Reliability for Speaker Recognition

Figure 2 illustrates that speaker recognition tasks are affected by expressive speech. When the emotional descriptors for arousal and valence are between 3 and 5, the EER is very small, suggesting that for neutral speech, the speaker recognition system has a reliable performance. For example, Figure 3 shows all the bins with EER less than 2%, after smoothing the binary mask by applying erosion and dilation operations. We can approximate this reliable region with the shaded gray region. This observation leads us to define regions of reliability in the arousal-valence space.

We aim to define reliable and unreliable regions for speaker recognition tasks. Following the analysis in Figure 3, we define the reliable region with the rectangle centered at the origin illustrated with the shaded gray region in Figure 4(a). We expect that sentences in this region will have EER less than 2%. For the unreliable regions, we consider all the bins where the EER was greater than 3%. The shaded gray areas in Figure 4(c) illustrates the unreliable regions. The region between the 2% and 3% EER is referred to as the region of uncertainty, where we cannot make a decision about the reliability of the sentences. Figure 4(b) illustrates the rectangular boundaries in the arousal-valence space for uncertain regions. Notice that we arbitrarily set these thresholds in the definition of reliable, uncertain and unreliable regions. As we increase the number of speakers, change the speaker recognition system, or increase the diversity of our sentences, we expect a different speaker recognition performance. However, these threshold seems appropriate for this study when we consider the results in Figure 2.

4.2. Predicting Reliability with Emotion Classification

Having defined regions of reliability for the speaker recognition task, the challenge now is to automatically predict where the test speech belong. We formulate the reliability estimation as a three class problem following the regions in Figure 4. Using acoustic features, we train a deep learning classifier where the goal is to determine if a speech sample is on the reliable, uncertain or unreliable region. In addition to accuracy in this multiclass problem, we are also interested in determining the EER for sentences predicted in each class, where the goal is to obtain similar numbers as the one obtained with the ground truth emotional labels.

4.2.1. Acoustic features. The study employs the popular feature set introduced for the Interspeech 2013 Computational Paralinguistic Challenge. We first extract a set of frame level features referred to as *low-level descriptors* (LLDs). The LLDs include fundamental frequency, MFCCs, zero crossing rate among other features. For each sentence, a set of global statistics such as arithmetic mean and standard deviation are calculated over the LLDs, which are referred to as *high-level features* (HLFs). The IS2013 feature set contain 6,373 HLFs per utterance. More details on the feature set can be found in [25]. The features are extracted using the OpenSmile toolkit [26].

4.2.2. Classification framework. We use a *deep neural network* (DNN) architecture to classify the test utterances into three classes. The speaker recognition task include data from 80 speakers (7,796 sentences). Since the speaker recognition results are only available for the 5,818 sentences used to test the speaker recognition models, we use the same testing set for the emotion recognition task. The DNN has parameters that are optimized using the validation set consisting of 15 speakers (790 utterances). To facilitate that our classification models are trained with speaker independent partitions, we use only the data from the rest of the corpus (4,846 utterances) to train our classifiers. Figure 1

TABLE 2. CONFUSION MATRIX BETWEEN THE CORRECT AND PREDICTED CLASSES FOR THE EMOTIONAL CLASSIFIERS

		Predicted Class		
		Reliable	Uncertain	Unreliable
Correct Class	Reliable	905	822	136
	Uncertain	973	1398	426
	Unreliable	201	540	417

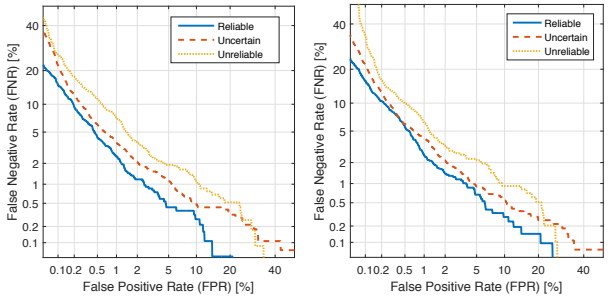
summarizes the partitions used for the emotion recognition task.

The proposed DNN consists of two hidden layers with 1,024 nodes, implemented with *rectified linear unit* (ReLU). A dropout of 0.5 is used at the input layer (features) and the first hidden layer to prevent overfitting. The loss function for the DNN is the cross entropy between the true class label and predicted label. The network is trained with 100 epochs with early stopping based on performance on the validation set. We use the z-normalization technique to normalize the input features, where we subtract their mean, and divide them by their corresponding standard deviation. The mean and standard deviation values are calculated from the sentences in the training set. After normalization, we realize that there are some unreliable features whose deviation from the mean is quite large (> 3). If the features are normal distributed, only 1% of them should fall outside this range. Notice that the data is very heterogeneous, with different recording conditions. We reduce the effect of these features by setting their value to zero after z-normalization.

4.3. Results

We evaluate the performance of the multi-class emotion recognition problem. Table 2 gives the confusion matrix for the original and predicted classes. From the confusion matrix we evaluate the F1-score for the three classes. The F1-score is 0.46 for the reliable class, 0.5 for the uncertain class and 0.39 for the unreliable class. Notice that assigning random classes would give a F1-score of 0.33. Therefore, our classifier performs above chances. This is a nonconventional speech emotion recognition problem where the classes have irregular boundaries in the arousal-valence space. In spite of the challenges of this classification problem, our system is able to obtain an average F1-score of 0.45.

More important than the speech emotion recognition results is the speaker recognition performance for sentences identified as reliable, uncertain and unreliable. First, we calculate the EER when the classes are defined using the ground-truth emotional labels. The results are estimated over all the sentences that fall within the boundaries of the respective classes. Figure 5(a) reports the *detection error tradeoff* (DET) curves for the speaker recognition task for each of the three classes. We observe a clear separation between the reliable and unreliable classes. The EER can be calculated as the point when the *false positive rate* (FPR) equals the *false negative rate* (FNR). The first row in Table 3 shows the EER for the three classes defined with the ground truth labels. The EER for the unreliable class (2.85%) is almost



(a) Ground truth emotional labels (b) Predicted classes

Figure 5. DET curves for speaker recognition task using reliability classes from the (a) emotional ground truth labels, and (b) predicted classes.

TABLE 3. EER FOR THE SPEAKER RECOGNITION TASKS FOR DIFFERENT REGIONS OF RELIABILITY. ROWS INDICATE THE TYPE OF TASK, ORIGINAL, PREDICTED CLASSES

Type	EER Reliable	EER Uncertain	EER Unreliable
Original	1.47	2.03	2.85
Predicted	1.64	2.04	2.76

double the EER for the reliable class (1.47%). Ideally, the EER for the predicted classes will maintain this separation.

Figure 5(b) shows the DET curves when the classes are predicted with the classifier. We follow the same procedure, where the speaker recognition results are estimated over all the sentences assigned to each of the three classes. The figure shows similar results as the ones obtained with the ground truth emotional labels (Fig. 5(a)). The second row in Table 3 shows the EER for the predicted classes. The table shows that the EER for both conditions are very similar. This analysis validates the methodology to assess speaker reliability based on emotional speech classifiers.

5. Conclusions

This paper proposed to predict the reliability of a speaker recognition task by considering the emotional content of the sentence. We presented a comprehensive analysis from 80 speakers to understand the performance of a speaker recognition system as a function of arousal and valence scores. We created a mismatch by training the speaker models with neutral speech and testing it with expressive speech. The analysis showed that emotional speech indeed affected the speaker recognition performance, especially for extreme values of arousal and valence. The analysis provided regions in the arousal-valence space for which we expect to have reliable speaker recognition results. This observation motivated us to train a speech emotion classifier to identify sentences belonging to reliable, uncertain and unreliable classes. We formulated this problem as a three class problem, training our speech emotion classifier to predict the reliability of a given sentence. The evaluation demonstrated that the DET curves and EER values are similar when the reliable, uncertain and unreliable classes

are defined either with ground truth emotional labels or with the predictions of our classifier.

This study demonstrated the potential of using emotion recognition in speaker recognition tasks. There are several directions that we are planning to explore. First, we will work on extending the MSP-PODCAST corpus including more speakers and diversifying the emotional content of the corpus. Second, we will concentrate on training better models for the emotion recognition task. We will tackle the problem using regression models that predict the values for arousal and valence. Our previous work has shown the benefits of performing regression by jointly learning the emotional dimensions with a multi-task architecture [27]. We will extend this study using those techniques to improve the robustness in predicting the reliability of a speaker recognition task. Third, we will estimate the relation between EER and emotional attributes (i.e., Fig. 2) by evaluating multiple speaker recognition/verification frameworks, not just one as proposed in this study. This extension will generate a smoother mapping that is more general. Fourth, we will consider other factors affecting speaker recognition systems (e.g., channel). Finally, for sentences that are identified as unreliable, we will explore compensation schemes to improve the speaker recognition task. This approach will not affect the performance of sentences in the reliable group, which is an important advantage of the proposed solution.

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] I. Shahin, "Speaker identification in emotional talking environments based on CSPHMM2s," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1652–1659, August 2013.

[3] H. Bao, M. Xu, and T. Zheng, "Emotion attribute projection for speaker recognition on emotional speech," in *Interspeech 2007*, Antwerp, Belgium, August 2007, pp. 758–761.

[4] D. Li, Y. Yang, Z. Wu, and T. Wu, "Emotion-state conversion for speaker recognition," in *Affective Computing and Intelligent Interaction (ACII 2005)*, ser. Lecture Notes in Computer Science, J. Tao, T. Tan, and R. Picard, Eds. Beijing, China: Springer Berlin Heidelberg, October 2005, vol. 3784, pp. 403–410.

[5] M. V. Ghiurcau, C. Rusu, and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4944–4947.

[6] W. Wu, T. Zheng, M. Xu, and H. Bao, "Study on speaker verification on emotional speech," in *International Conference on Spoken Language (ICSLP 2006)*, Pittsburgh, PA, USA, September 2006, pp. 2102–2105.

[7] S. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, June 2012.

[8] Z. Wu, D. Li, and Y. Yang, "Rules based feature modification for affective speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, Toulouse, France, May 2006, pp. 661–664.

[9] S. Krothapalli, J. Yadav, S. Sarkar, S. Koolagudi, and A. Vuppala, "Neural network based feature transformation for emotion independent speaker identification," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 335–349, September 2012.

[10] I. Shahin, "Speaker identification in emotional talking environments using both gender and emotion cues," in *International Conference on Communications, Signal Processing, and their Applications (ICCSPA 2013)*, Sharjah, United Arab Emirates, February 2013, pp. 1–6.

[11] D. Li and Y. Yang, "Emotional speech clustering based robust speaker recognition system," in *International Congress on Image and Signal Processing (CISP 2009)*, Tianjin, China, October 2009, pp. 1–5.

[12] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2017.

[13] S. Parthasarathy, C. Zhang, J. Hansen, and C. Busso, "A study of speaker verification performance with expressive speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5540–5544.

[14] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[15] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.

[16] M. Huggins and J. Grieco, "Confidence metrics for speaker identification," in *International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, CO, USA, September 2002, pp. 1381–1384.

[17] W. Campbell, D. Reynolds, J. Campbell, and K. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 473–476.

[18] J. Richiardi, P. Prodanov, and A. Drygajlo, "A probabilistic measure of modality reliability in speaker verification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 709–712.

[19] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions," *Speech Communication*, vol. 78, pp. 42–61, April 2016.

[20] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[21] 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, May 2011.

[22] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2854–2857.

[23] C. Zhang, G. Liu, C. Yu, and J. H. Hansen, "I-vector based physical task stress detection with different fusion strategies," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2689–2693.

[24] G. Liu, T. Hasan, H. Boril, and J. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 7755–7759.

[25] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*, ser. Springer Theses. Springer, June 2017.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[27] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.