2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

# Exploring the Intersection Between Speaker Verification and Emotion Recognition

Michelle Bancroft, Reza Lotfian, John Hansen, Carlos Busso
*Multimodal Signal Processing (MSP) laboratory, Department of Electrical and Computer Engineering*
The University of Texas at Dallas, Richardson TX 75080, USA
mrb150630@utdallas.edu, reza.lotfian@utdallas.edu, john.hansen@utdallas.edu, busso@utdallas.edu

***Abstract*—Many scenarios in practical applications require the use of speaker verification systems using audio with high emotional content (e.g., calls from 911, forensic analysis of threatening recordings). For these cases, it is important to explore the intersection between speaker and emotion recognition tasks. A key challenge to address this problem is the lack of resources, since current emotional databases are commonly limited in size and number of speakers. This paper (1) creates the infrastructure to study this challenging problems, and (2) presents an exploratory analysis to evaluate the accuracy of state-of-the-art speaker and emotion recognition systems to automatically retrieve specific emotional behaviors from target speakers. We collected a pool of sentences from multiple speakers (132,930 segments), where some of these speaking turns belong to 146 speakers in the MSP-Podcast database. Our framework trains speaking verification models, which are used to retrieve candidate speaking turns from the pool of sentences. The emotional content in these sentences are detected using state-of-the-art emotion recognition algorithms. The experimental evaluation provides promising results, where most of the retrieved sentences belong to the target speakers and has the target emotion. The results highlight the need for emotional compensation in speaker recognition systems, especially if these models are intended for commercial applications.**

***Index Terms*—Speech emotion recognition, speaker verification, computational paralinguistics.**

## I. Introduction

Speaker verification tasks are often conducted using emotional speech. Examples includes emergency calls, forensic analysis, and surveillance recordings. The key challenge in this area is that the performance of speaker verification systems is affected by emotional speech [1], [2]. Unfortunately, the progress on speaker verification in the presence of emotion has been limited due to the lack of appropriate databases, which are small with only a few subjects. This limitation has clear implications in the deployment of speaker verification system in real applications. It is important to explore the intersection between speaker recognition and emotion recognition tasks. Toward this goal, our group is interested in exploring the problem of retrieving sentences with target emotion spoken by target individuals (e.g., detecting highly aroused recordings from "Joe" from a large audio repository).

The main focus of this study is exploring the feasibility of retrieving from a large audio repository recordings from target individuals conveying given emotional behaviors using existing state-of-the-art speaker and emotion recognition systems. The first contribution of this study is to create a unique infrastructure to address this problem. An ideal database for this retrieval task needs to have emotional speech from multiple speakers, where each speaker has enough recordings to build robust speaker verification models. We also need a large audio repository with recordings from the target subjects to retrieve the target speech segments. We create this infrastructure by downloading podcasts from audio-sharing websites conveying a variety of emotional content. The podcasts are segmented into speaking turns, where some of them are annotated with emotional labels. This is an ongoing project, where we have annotated with emotional labels data from 146 speakers, each of them with over 150 seconds of audio. We also have 132,930 unlabeled segments, which serve as our audio repository. Some of the speaking turns in this unlabeled pool of sentences belong to the 146 target speakers. This infrastructure provides the ideal resource to explore the intersection between speaker and emotion recognition tasks.

The second contribution of the study is an exploratory analysis to evaluate the performance of current state-of-the-art speaker and emotion recognition systems to retrieve expressive speech from target individuals. The proposed approach consists of recognizing speech segments belonging to target individuals using a speaker verification system, which are then emotionally evaluated, using an emotion recognition system. We build speaker verification models using the i-vector framework with *probabilistic linear discriminant analysis* (PLDA) as a back-end. After training speaking models, we automatically identify speaking turns from the target speakers that convey the target emotional content. We rely on state-of-the-art emotion recognition algorithms to predict arousal and valence scores, using multitask deep learning architectures similar to the models proposed by Parthasarathy and Busso [3]. To evaluate the performance of the proposed system, we manually annotate the speaker identity of the retrieved samples. These segments are emotionally annotated with crowdsourcing. From the speech repository, we retrieved 1,003 samples that our model predict to belong to 146 target speakers conveying target emotions (four corners in the arousal-valence space). Even without compensating for emotional content, over 80% of the retrieved sentences belong to the target speakers. 45.8% of the retrieved samples belong to the target emotional regions,

This work was funded by NSF CAREER award IIS-1453781.

978-1-7281-3891-6/19/$31.00 ©2019 IEEE

and 77.4% belong to their target quadrant. These encouraging results validate this new novel formulation to combine speaker and emotion recognition systems, creating the foundation to improve speaker recognition tasks in the presence of emotional speech.

## II. Related Work

Most studies exploring speaker verification tasks using emotional speech have focused on quantifying the drop in performance caused when these systems are used with emotional speech. Other studies have proposed compensation schemes to mitigate the drop in performance for emotional speech.

Staroniewicz [4] analyzed the negative effect of categorical emotions on the performance of a speaker identification system with *Mel-frequency cepstral coefficients* (MFCCs) as features, and with a back-end implemented with *Gaussian mixture models* (GMMs). The study concluded that the performance degrades more for emotions with high arousal. Parthasarathy and Busso [1] explored the degradation in the performance of speaker recognition tasks due to the mismatch in the emotional content between the train and test sets. Their analysis showed the regions in the arousal and valence space where a speaker verification system can be considered reliable. These analyses confirmed the findings reported in early studies on speaker verification in the presence of emotion [5]–[9].

The negative effects of emotion on speaker verification tasks have inspired researchers to develop speaker verification systems that are less affected by these variations [6], [10]–[13]. A solution is to explore features that are robust in discriminating speaker information, but are less sensitive to emotional changes. Krothapalli et al. [10] proposed emotion-dependent feature transformations with neural networks to compensate for emotional variations. Li et al. [11] proposed an emotion-state conversion approach to improve the performance of speaker identification system when tested with emotional speech. An alternative solution is to change the back-end of the system. Shahin [13] proposed to use a second-order *circular suprasegmental hidden Markov models* (CSPHMM2s) as a classifier. The study used log frequency power coefficients, showing improvements over other speaker verification systems. Wu et al. [6] proposed to normalize the speaker verification scores according to the emotion.

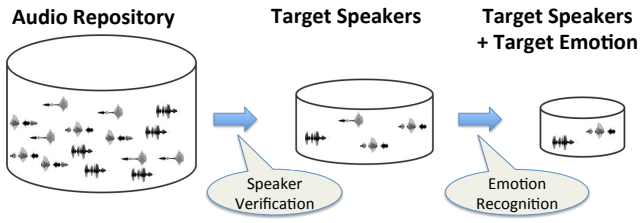



Fig. 1. Proposed analysis to evaluate existing speaker and emotion recognition systems to retrieve speech segments from a target speaker conveying target emotion.

To the best of our knowledge, this is the first time that speaker and emotion recognition tasks are combined as a retrieval task where the goal is to identify speech samples with a given emotion, spoken by a target speaker. This is an important problem with clear applications in forensic analysis. This study is unique in that it creates the infrastructure to address this problem with (1) labeled data with speaker and emotional information to train speaker and emotion recognition systems, and (2) a large set of unlabeled data from where we retrieve the target sentences. While the building blocks for the emotion and speaker recognition systems rely on frameworks proposed by other studies, the contributions of this study are (1) building the infrastructure for this task, and (2) analyzing the feasibility of retrieving emotional speech from target individuals using current state-of-the-art speaker and emotion recognition systems.

## III. Infrastructure for the Study

An important contribution of this study is building the infrastructure to study speaker and emotion recognition tasks. We use the MSP-Podcast database [14], which is a corpus developed following the ideas presented in Mariooryad et al. [15]. The MSP-Podcast database is a collection of naturalistic, emotional data which comes from over 1,000 podcasts available at online audio sharing websites. Each podcast is segmented into speaking turns using a speaker diarization tool, where segments with music, multiple speakers and background noise are discarded. This automatic pipeline is not perfect and some of the selected segments have multiple speakers, music or noise. This is an ongoing effort, where we currently have 152,975 speech segments (we use version 1.0 of the corpus). We have annotated 20,032 speaking turns with emotional labels, using a modified version of the crowdsourcing method introduced by Burmania et al. [16]. This study uses attribute-based emotional descriptors for arousal (calm versus active), and valence (negative versus positive).

Out of the 20,032 speaking turns with emotional labels, we have manually annotated the speaker identity of 16,015 segments. We have 146 speakers with more than 150 seconds of recordings, providing enough data to train robust speaker verification models. These 146 speakers define our *target* speakers. Segments without emotional labels (i.e., 132,930 speaking turns) forms our audio repository from which we will retrieve speech segments. These unlabeled speaking turns are not part of the version 1.0 of the MSP-Podcast corpus. However, several of the speakers appear in multiple podcasts, so the audio repository has segments spoken by our 146 target speakers. This infrastructure is ideal to explore the intersection between speaker and emotion recognition systems. It has not only labeled data with both emotion and speaker information, but also a large unlabeled speech repository for retrieval tasks.

## IV. Formulation for the Exploratory Analysis

This section describes the novel formulation proposed to evaluate the feasibility of using existing technologies for speaker and emotion recognition to retrieve emotional speech

from target speakers. Figure 1 describes the pipeline for our analysis. The two-stage approach identifies the segments from the target speakers, and predicts the emotional content of these segments. The retrieved speaking turns are the segments that satisfies the required constraints. This section describes the building block of our system.

### A. Speaker Verification System

The first step in the proposed analysis is to identify sentences from the target speakers in the audio repository. This problem is formulated as a speaker verification task where each segment is compared with the 146 speaker models. We rely on the i-vector framework with *probabilistic linear discriminant analysis* (PLDA) as a back-end [17]. We extract MFCCs with their $\Delta+\Delta\Delta$ features, creating a 39-dimensional vector. Then, we create the i-vector with:

$$M = m + T\mathbf{x} \tag{1}$$

where $M$ is a GMM supervector obtained with *maximum a posteriori* (MAP) adaptation of a *universal background model* (UBM). This vector is written as the summation of two terms. The first term is the vector $m$, which is the mean-vector, independent of the channel, emotion and speaker variability. The second term is the product $T\mathbf{x}$. $\mathbf{x}$ is the i-vector, which is a low dimensional vector that multiplies the total variability matrix $T$. The i-vectors are normalized and used as input for our back-end implemented with *probabilistic linear discriminant analysis* (PLDA) [18]. During the back-end process, the speaker models are created and the likelihood for each segment belonging to a speaker is scored. During the evaluation, the test segments are also represented as normalized i-vectors. The *log-likelihood ratio* (LLR) is then calculated for each segment using Equation 2. $\mathbf{x_1}$ represents the normalized i-vector for the enrolled speaker and $\mathbf{x_2}$ is the normalized i-vector for the test speech segment. The LLR computes the ratio between two alternative hypothesis: $H_1$ : $\mathbf{x_1}$ and $\mathbf{x_2}$ come from the same speaker model, and $H_0$: $\mathbf{x_1}$ and $\mathbf{x_2}$ come from different speaker models. The vector $\mathbf{x_1}$ and $\mathbf{x_2}$ are modeled using the Gaussian PLDA. The higher the score, the more confident the system is that a segment belongs to a specific speaker.

$$r = \ln \frac{\rho\left(x_1, x_2|H_1\right)}{\rho\left(x_1|H_0\right) \cdot \rho\left(x_2|H_0\right)} \tag{2}$$

### B. Emotion Recognition System

The second step in our analysis is to detect the emotional content on the speech segments that passed the speaker identification process, retrieving sentences conveying the target emotion. An important step is to define the emotional content of interest, which clearly depends on the application. This study describes emotion with attribute-based annotations. The evaluators used a 7-point Likert scale for valence (1-very negative, 7-very positive), and arousal (1-very calm, 7-very active). Arousal and valence are the most common emotional attributes, defining a convenient space for our analysis. The regions of interest for this study are the four extreme corners of the arousal-valence space (see Fig. 4):

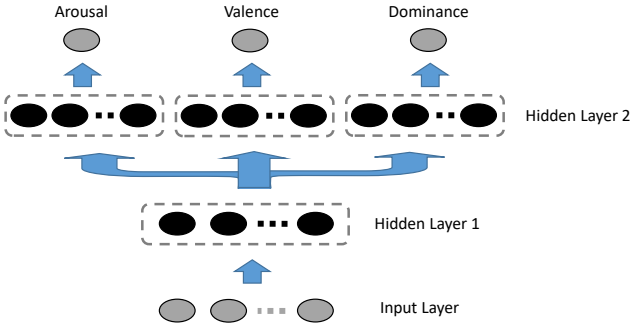



Fig. 2. DNN architecture for MTL framework. The first hidden layer is shared across attributes. The second hidden layer is specific to each attribute.

- Region 1: low valence, high arousal [Val.$\in$ 1-3;Aro.$\in$ 5-7]
- Region 2: high valence, high arousal [Val.$\in$ 5-7;Aro.$\in$ 5-7]
- Region 3: low valence, low arousal [Val.$\in$ 1-3;Aro.$\in$ 1-3]
- Region 4: high valence, low arousal [Val.$\in$ 5-7;Aro.$\in$ 1-3]

Region 1 includes negative emotions such as anger and fear. Region 2 includes positive emotions such as happiness and excitement. Region 3 includes negative emotions with low arousal such as sadness. Region 4 includes positive emotions with low arousal such as contempt. The center of the arousal-valence space is [4,4] which includes mostly neutral sentences.

The emotional detection task is to predict the arousal and valence scores of a speech segment using acoustic features. There are many frameworks that can be used for this task [19]–[21]. We rely on a recent approach proposed by Parthasarathy and Busso [3], which uses *multitask learning* (MTL) to jointly predict the arousal, valence and dominance scores. The approach relies on the *deep neural networks* (DNNs) presented in Figure 2 where the cost function considers the error in predicting the three emotional attributes. In this study, we re-implement this framework following the configuration proposed in that study, training the models with a portion of our recordings annotated with emotional labels. The training set has 6,710 segments and the development set has 887 segments. We train a model for arousal, where the weights associated with the prediction errors for arousal, valence and dominance are adjusted to maximize the performance for arousal on the development set. A similar network is built for our valence predictor. These models are trained with the feature set proposed for the computational paralinguistic challenge in Interspeech 2013 [22], which consists of 6,373 features.

The MTL structure is built with *rectified linear unit* (ReLU), dropout ($p$=0.5), stochastic gradient decent with learning rate of $1e^{-4}$ per sample, mini-batch size of 256 and a constant momentum of 0.9, following the description in Parthasarathy and Busso [3].

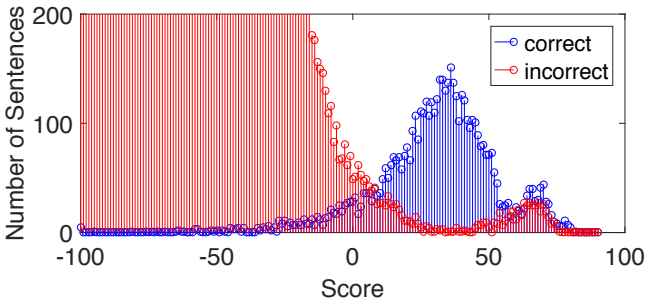



Fig. 3. Histogram of the results speaker verification results for the annotated data (Eq. 2). The blue bars correspond to the segments correctly identified by the speaker verification system. The red bars represent the segments incorrectly identified by the system as function of the score. The histogram is cropped for better visualization of the correct samples.

## V. Experimental Evaluation

This section describes the results of the experimental evaluation of the proposed analysis to retrieve sentences from target speakers conveying target emotions.

### A. Speaker Verification Task

We conduct the speaker verification test for each of the target speakers on the speech segments from the audio repository. The result of the test is the log-likelihood ratio in Equation 2. The higher the ratio, the higher the confidence of the speaker verification system in assigning the speaker identity to a given sentence. An important step is to set a reasonable threshold for this ratio. To address this problem, we trained the speaker verification models with only 150 seconds, leaving the remaining data for validation. Then, we evaluated the speaker verification models on this validation set, comparing each speaker model to each sentence. Figure 3 shows the histogram of the ratio for positive cases (speaker model corresponds to the actual speaker – blue histogram) and negative cases (speaker model does not correspond to the actual speaker – red histogram). The histogram for negative cases is cropped to increase the resolution around reasonable values for $r$. The figure shows that a reasonable threshold is $r = 10$. We take a more conservative threshold equals to $r = 12$, aiming to increase the precision rate.

The speaker verification models identified 33,628 unique segments from the audio repository with a score greater than $r = 12$. These sentences are analyzed by the emotion prediction systems.

### B. Emotion Recognition Task

We evaluate the emotion prediction systems for arousal and valence on each of the sentences retrieved from the target subjects (33,628 speech segments). Figure 4 shows the dispersion of these sentences in the arousal-valence space using the predicted values. The figure also shows the four target regions. There was a total of 1,003 unique segments in the target regions. The majority of the segments belonged to regions 1 and 2 (region 1: 294 region 2: 681; region 3: 15; region 4: 13).

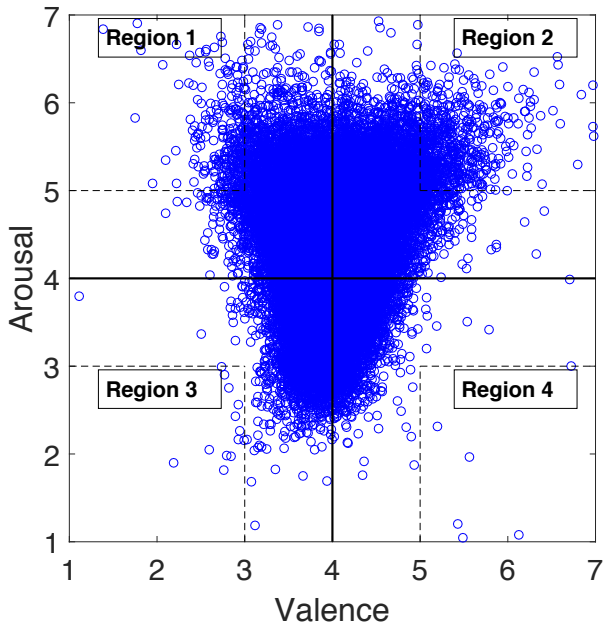



Fig. 4. Results from the emotion recognition system on the 33,628 sentences retrieved by the speaker verification system. Sentences in the four corners in the arousal-valence space are selected and emotionally annotated.

### C. Analysis of the Results

This section further analyzes the 1,003 speaking turns that satisfy both conditions (target speaker and target emotional region). The speaker verification tests are independently conducted for each speaker model. Therefore, a segment can be assigned to more than one speaker as long as $r$ in Equation 2 is greater than 12. A sentence can have more than one speaker, as the conversations are spontaneous with overlapped speech. From the 1,003 unique sentences, there are 1,401 speaker verification evaluations that satisfy this ratio.

We manually annotate the speaker identity of these segments. This process is conducted by listening to the segments. We define a conservative approach, where the following criteria have to be reached to consider the speaker label correct: (1) speaker sounds similar to the voice in the segment, (2) the segment belongs to a podcast containing annotated segments from the speaker, (3) speaker is active in the surrounding segments of the podcast. Notice that some of the segments have overlapped speech resulting in multiple correct/wrong answer per file.

Out of 1,401 evaluations, the speaker verification system successfully identified the speaker information of 1,135 evaluations with an accuracy of 80.9% (a sentence can have more than one speaker, where the total number of cases is 1,401). Notice that we expected better speaker verification performance if the target regions include neutral areas within the arousal-valence space. Given that the retrieved segments are predicted to convey emotions (regions 1-4), this level of accuracy is expected.

The retrieved samples are annotated with emotional labels

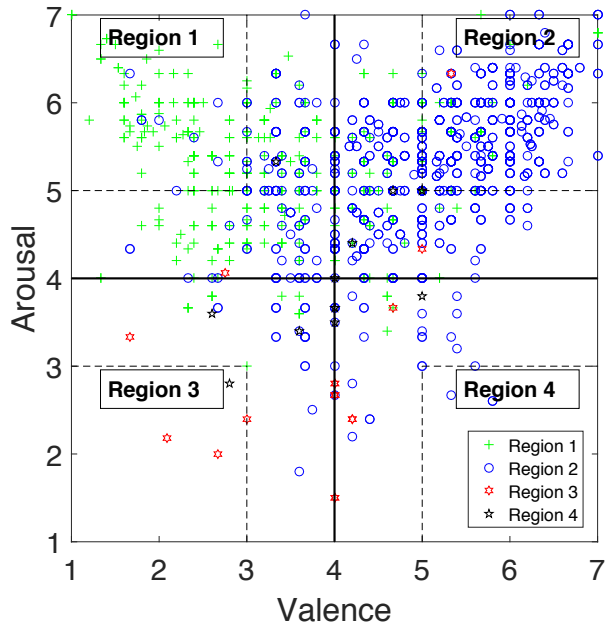
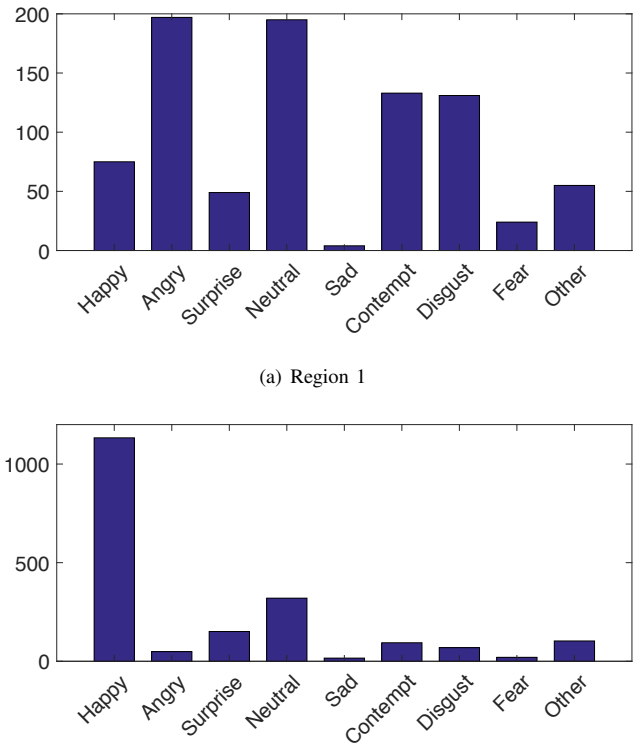



Fig. 5. Emotional content of the retrieved segments displayed on the arousal-valence space (average results of the emotional annotations). Most segments belong to their corresponding quadrants.




(a) Region 1

(b) Region 2

Fig. 6. Individual emotional labels assigned to the retrieved segments (each file is independently annotated by three speakers). The figure presents the results for regions 1 and 2.

to evaluate the performance of the emotion detection system. For the subjective evaluations, we rely on a crowdsourcing platform using *Amazon Mechanical Turk* (AMT). Before uploading samples for subjective evaluations, we listened to the segments. We discarded 35 segments from the 1,003 retrieved segments since they have either only music, or mostly silence (i.e., 968 speaking turns are emotionally annotated). We implement the same protocol and questionnaire used to emotionally annotate our labeled set (20,032 sentences described in Sec. III). Every three samples, we randomly add one sentence from the training corpus with known labels as a reference to monitor the performance of the workers during the evaluation. We stop the perceptual evaluation when the performance is below an acceptable threshold, following the approach introduced by Burmania et al. [16]. We collected three evaluations per speech segment.

Figure 5 shows the average arousal and valence values in the annotations of the retrieved sentences. The figure highlights the four target regions, where we expected these sentences to belong. If we consider the target regions, the precision rate for sentences in the actual regions are 37.5% for region 1, 50.6% for region 2, 23.1% for region 3 and 0% for region 4 (45.8% overall precision rate). Notice that there are few sentences for regions 3 and 4, so the performance need to be considered with caution. If we consider the quadrants instead of the actual regions, the precision rates increases to 73.3% for region 1, 80.2% for region 2, 61.5% for region 3, and 36.4% for region 4 (77.4% overall precision rate). We also estimate the *concordance correlation coefficient* (CCC) between the predicted and annotated scores for arousal and valence. We obtain $ccc_{aro.}$=0.532 for arousal, and $ccc_{val.}$=0.364 for valence. While the performance for valence is higher than the CCC reported in previous studies [3], the performance for arousal is lower than expected.

We also evaluate the histogram of the emotional categories retrieved for each region. We include all the individual annotations assigned to the retrieved segments (i.e., three annotations per speech turn). Since there are few sentences for regions 3 and 4, we focus the analysis on regions 1 and 2, which are shown in Figure 6. Figure 6(a) shows that the retrieved sentences in region 1 are mostly neutral, anger, contempt and disgust. We expected negative emotions in this region. Figure 6(b) shows that the retrieved sentences for region 2 are mostly happy, as expected (see analysis in Busso and Narayanan [23]).

Finally, we analyze the speaker verification performance as a function of the target emotional regions. The overall error of the speaker verification system using $r = 12$ is 19.1%. The average error per regions are: 18.4% (region 1), 19.1% (region 2), 20.0% (region 3), and 33.3% (region 4). With the exception of region 4, which has only 13 sentences, the performance is consistent across regions. Notice that the retrieved samples are expected to be emotional. We expect better speaker verification results on emotionally neutral sentences.

## VI. Conclusions

This study made two important contributions. First, we created the infrastructure to explore retrieval problems for speaker recognition systems in the presence of emotion. The setting included labeled data with emotion and speaker information (i.e., MSP-Podcast corpus), and a large speech repository with unlabeled data that included segments from the target speakers. Second, we evaluated the feasibility of using existing state-of-the-art techniques for speaker and emotion recognitions to retrieve emotional data from specific speakers. Using a speech repository of 132,930 speech segments, we retrieved 1,003 sentences that satisfy both conditions: they are expected to belong to the 146 target speakers, and they are expected to conveyed the target emotion (regions 1 to 4 in Fig. 4). To measure the performance of our proposed pipeline, we annotated the emotional content and the speaker identity of the retrieved speech segments. The results showed that over 80.9% of the speaker verification evaluations were correct. The results also showed adequate performance for emotion recognition tasks, where 45.8% of the turns belong to the target region in the arousal-valence space, and 77.4% belong to its target quadrant.

This study builds the foundation for future research in emotional retrieval of speech segments spoken by target speakers. This paper uses predefined areas in the arousal-valence space as the target regions. We can also formulate the problem as a retrieval of categorical classes (e.g., retrieval of angry sentences from "Joe"). The combination of emotional recognition and speaker verification techniques proposed in this paper can lead to important tools for forensic analysis (i.e., resources to retrieve threatening behaviors from target individuals from large audio repository). The results reported in this study depends on the performance of the speaker and emotion recognition systems. We expect improved performance in the proposed retrieval system by improving individual systems. The analysis revealed limitations of speaker verification tasks in the presence of emotional speech. We expect that novel speaker verification techniques that compensate for emotional variations can lead to better performance (e.g., feature normalization, extraction of robust features, robust speaker verification models).

## References


[1] 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.

[2] S. Parthasarathy and C. Busso, "Predicting speaker recognition reliability by considering emotional content," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 434–436.

[3] ——, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[4] P. Staroniewicz, "Considering basic emotional state information in speaker verification," in *International Workshop on Biometrics and Forensics (IWBF 2016)*, Limassol, Cyprus, March 2016, pp. 1–4.

[5] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Interspeech 2008*, Brisbane, Australia, September 2008, pp. 609–612.

[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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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, 2019.

[15] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[16] 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.

[17] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[18] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.

[19] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[20] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[21] ——, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088.

[22] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[23] C. Busso and S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 257–260.