INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English

Jacob Sager, Ravi Shankar, Jacob Reinhold, Archana Venkataraman

Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore MD 21218, USA
{jsager, rshanka3, jacob.reinhold, archana.venkataraman}@jhu.edu

Abstract

We introduce the Varied Emotion in Syntactically Uniform Speech (VESUS) repository as a new resource for the speech community. VESUS is a lexically controlled database, in which a semantically neutral script is portrayed with different emotional inflections. In total, VESUS contains over 250 distinct phrases, each read by ten actors in five emotional states. We use crowd sourcing to obtain ten human ratings for the perceived emotional content of each utterance. Our unique database construction enables a multitude of scientific and technical explorations. To jumpstart this effort, we provide benchmark performance on three distinct emotion recognition tasks using VESUS: longitudinal speaker analysis, extrapolating across syntactical complexity, and generalization to a new speaker.

Index Terms: Emotional speech dataset, lexical consistency, perceptual variability, emotion recognition benchmark

1. Introduction

Emotion is the cornerstone of human social interactions and is becoming increasingly relevant in speech processing applications. For example, automated emotion recognition is used in the transportation sector [1], in mental health evaluations [2], and in service-oriented environments [3]. Most recognition platforms are based on carefully selected features, such as the fundamental frequency (F_0), signal energy, articulation rate, and MFCCs [4, 5, 6]. However, the field is now migrating towards data-driven alternatives to feature selection using deep neural networks [7, 8, 9]. While current emotion recognition methods perform well on simple classification tasks, they rarely generalize across datasets [10]. Along the same lines, parametric synthesis models can produce emotional speech on a case-by-case basis [11, 12], but we have a limited understanding of what internal settings correspond to each emotional class.

One of the roadblocks to developing robust models for emotional speech is the scarcity of freely available training data, particularly for North American English. The current benchmark for emotion recognition is the IEMOCAP database collected at the University of Southern California [13]. IEMOCAP focuses on dyadic interactions between two actors. The database contains both scripted and improvised utterances across a range of emotionally-charged scenarios. IEMOCAP has become an invaluable resource to the speech community; however, it has two crucial limitations. First, due to the nature of dyadic improvisation, IEMOCAP has little control over lexical content. Said another way, there are very few examples of the same phrase being spoken with different emotional inflections. This prevents us from isolating the contributions of vocabulary, speaker identity, and syntactical complexity to emotional content. Second, the utterances are evaluated by only three expert raters. As such, the IEMOCAP database does not capture the variability of emotional perception across the general population [10, 14].

Prior work has recognized the need for a lexically controlled emotional speech database. Here, the relevant resources for North American English include the RAVDESS [15], SAVEE [16] and MSP-IMPROV [17] databases. While these databases contain a range of speakers and emotional categories, they are limited in terms of vocabulary. At one extreme, RAVDESS is built around just *two neutral sentences*. The SAVEE and MSP-IMPROV databases are more diverse, with MSP-IMPROV using an advanced emotion elicitation technique. However, SAVEE contains just *three sentences* that are common across emotional classes, and MSP-IMPROV includes only *fifteen target sentences* that are spoken with multiple emotional inflections. While valuable, the restricted vocabulary in these databases may not be sufficient to learn a comprehensive and generalizable model for emotional speech.

This paper introduces the Varied Emotion in Syntactically Uniform Speech (VESUS) repository, which fills an unmet need in currently available emotional speech resources. Unlike prior work, we build syntactical complexity from single- and multi-word phrases to complete sentences. Our repository consists of 252 unique phrases spoken by 10 actors in five hallmark emotional categories: happiness, sadness, anger, fear, and neutral. Our human evaluation consists of a large crowd-sourcing experiment, in which we obtain categorical labels from 10 raters. This granularity allows us to quantify the perceptual variability of each utterance. As demonstrated in Section 4, these ratings can also be used to moderate the difficulty of emotion recognition tasks. In total, the VESUS repository contains over 12,000 speech utterances and over 120,000 emotional annotations. Hence, it is a natural complement to existing resources and will enable new explorations by the general research community. Our audio recordings and crowd-sourced annotations will be made publicly available for download at `https://engineering.jhu.edu/nsa/`.

2. Database Design and Acquisition

VESUS focuses on semantic variety through scripted emotions and perceptual variability across individuals via crowd sourcing. Our repository will provide a much-needed stepping stone for generative and discriminative models of emotional speech.

2.1. Lexically Diverse Script

We consider several factors when designing our VESUS script. Semantic information is one obvious component. For example, the statement "I got a promotion" is more likely to convey positive sentiment than the statement "I failed the midterm," irrespective of the vocal inflections. Likewise, longer phrases pro-

http://dx.doi.org/10.21437/Interspeech.2019-1413

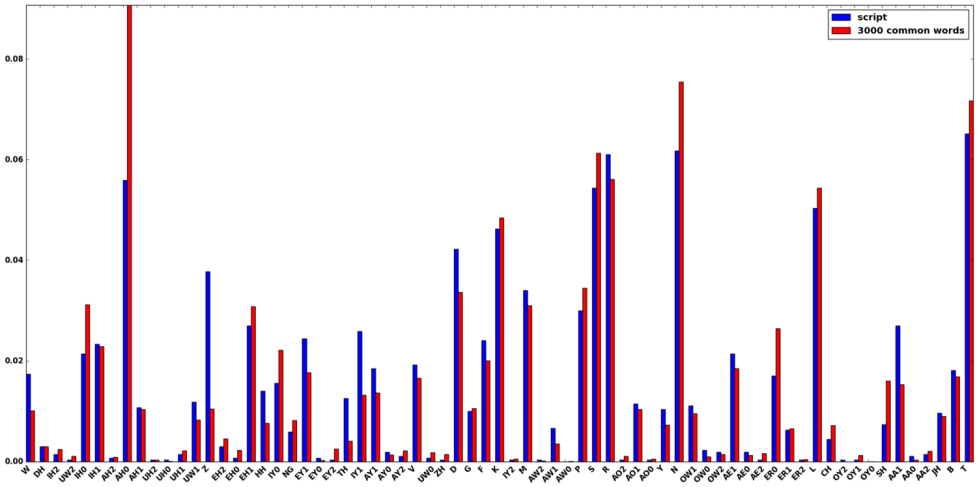

Figure 1: *Phonetic comparison of our script (blue) with the 3,000 most commonly used words in spoken English (red). Bins corresponds to one of the 44 English phonemes. The y-axis indicates the frequency of occurrence. The VESUS script is phonetically balanced.*

vide context to stress and intonation, but they allow semantics to play a greater role in the emotional perception [18].

Given these considerations, we have chosen to focus on low-level syntaxes to link emotional speech with the building blocks of language. Our VESUS script contains semantically neutral single- and multi-word phrases across a range of topics. Single-word categories include colors, numbers and geographic features. Multi-word phrases consist of dates and larger numbers. At the sentence level, we consider the $\langle noun, verb, predicate \rangle$ structure as a fixed unit of syntactical complexity. Our script is built around one and two syllable words, which is representative of North American English. However, we have included phrases with 3-5 syllable words to probe an additional layer of complexity. Finally, we have incorporated 15 emotionally-charged sentences to evaluate the role of semantics in the production and perception of emotional speech.

Table 1: *Top: Utterance level statistics for our VESUS script. Bottom: Acquisition information for the VESUS repository.*

Single Word Utterances	45
Multi-Word Phrases	207
Sentences: ⟨noun, verb, predicate⟩	159
Number of Scripted Utterances	**252**
VESUS Vocabulary (# Unique Words)	**449**
Average Duration per Utterance	1.74 sec
Average Duration per Actor	36 min
Average Duration per Emotion	73 min
VESUS Repository Size (Utterances)	**12,594**
VESUS Repository Duration	**6 h 9 min**

Fig. 1 verifies the phonetic balance of our script, as compared to the 3,000 most commonly used words in spoken English [19]. As seen, not only does our script sample each phoneme class, but the occurrence frequencies are closely aligned with real-world speech vocabulary. Table 1 (top) summarizes the utterance-level statistics of our script. For comparison, the RAVDESS database contains only *8 unique words*, and the target sentences in MSP-IMPROV *contain just 81 unique words*, including articles, prepositions and pronouns.

Table 2: *Crowd sourcing statistics for the VESUS repository.*

Number of Unique Participants	389
Average Ratings per Participant	372
Average Time per Rating	36 sec
Total Number of Emotional Ratings	125,940

2.2. Audio Recording Procedures

We recruited ten English speaking actors (5 male, 5 female) with varying professional experience from the Baltimore area. Informed consent was obtained prior to the session according to an approved IRB protocol. The audio recordings took place in a sound-proof environment on the Johns Hopkins University (JHU) campus. Our audio equipment consisted of an AKG pro audio C214 condenser microphone (cardioid) with adjustable stand, a Focusrite Scarlett 2i2 preamplifier, and GLS cords.

The actors received a paper copy of the script. They were first asked to read the entire script aloud in a neutral voice. This process was repeated for each of the following emotions: happiness, sadness, anger, and fear. The actors were instructed to pause between utterances to given themselves time to reset. They were also given a break between each script reading. Finally, we asked the actor to rate his/her level of confidence in each of their emotional portrayals (scale: 1–10).

2.3. Data Post-processing

We segmented the recordings on an utterance level using Audacity (v2.1.3), which relies on intensity-based thresholds. From here, we manually inspected each audio clip to ensure correct utterance boundaries. These post-processing steps eliminated both silence and unwanted conversational dialogue. In total, VESUS contains 12,594 unique utterances. Table 1 (bottom) summarizes the acquisition statistics of our VESUS repository. The durations correspond to the voiced segments. Our database will be released with an intuitive file hierarchy.

3. Large-Scale Human Evaluation

We have used Amazon Mechanical Turk (AMT) to crowd source ten emotional annotations for each of the 12,594 utter-

Table 3: *Confusion matrices for intended versus perceived emotion. Top: Individual rater accuracies. Bottom: Majority voting across individuals for each utterance (group-level annotations).*

Average Accuracy: 57%

		Intended Emotion				
		Neutral	Angry	Happy	Sad	Fear
Perceived	Neutral	**19687**	5749	9600	6016	6298
	Angry	1497	**16663**	2033	413	1782
	Happy	823	1378	**10312**	271	1204
	Sad	2736	564	1823	**15573**	6911
	Fearful	274	639	1217	2685	**8830**

Average Accuracy: 65%

		Intended Emotion				
		Neutral	Angry	Happy	Sad	Fear
Perceived	Neutral	**2354**	625	997	443	631
	Angry	40	**1805**	120	8	126
	Happy	3	51	**1143**	6	60
	Sad	121	11	148	**1848**	663
	Fearful	2	27	110	212	**1070**

ances. Our AMT task involves listening to a single recorded utterance and answering two simple questions. First, users were asked which emotion (happiness, sadness, anger, fear, neutral) best describes the attitude of the speaker. Users were explicitly instructed to base their decision on the tone of voice rather than on the semantic content. Second, users were asked to rate their level of confidence in their selected emotion (scale: 1–5). We did not query secondary emotional categories in this study to avoid influencing gut emotional reactions. We restricted our jobs to English-speaking AMT workers with masters level certification. We also monitored the quality of responses to ensure that workers were faithfully completing the task. Table 2 summarizes the crowd sourcing statistics of our VESUS repository. We emphasize that this is one of the largest crowd-sourcing endeavors in the emotional speech community.

Table 3 shows the confusion matrices for the intended versus perceived affect. We consider both the individual ratings (top) and the group-wise mode for each utterance (bottom), where ties are randomly assigned between the top categories.

We observe two important trends from these results. First, the individual accuracies are slightly lower than comparable audiovisual databases [13, 17]. One reason is that audiovisual raters have access to both voice and facial expressions. This combined information provides a significant advantage over judging emotions from audio recordings alone. In fact, our VESUS arguably provide a fairer baseline for emotional speech recognition algorithms. Another factor is that the VESUS repository is based on scripted rather than improvised utterances. We made this decision to maximize the size and diversity of our dataset given fixed resources. The second observation is that the accuracy jumps from 57% based on individual users to 65% for the group ratings. This trend can be observed within each emotional class and suggests a robust "wisdom of the crowd" phenomenon [20] in the speech regime.

Fig. 2 illustrates the average rater confidence across single words, multi-word phrases and complete sentences. We notice a slight decline in rater confidence for happiness and fear, which according to Table 3, are often confused with anger and sadness,

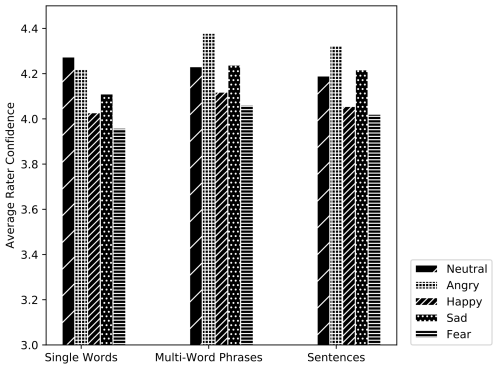

Figure 2: *Average confidence ratings across lexical category and emotional class.* ***1****:Not confident →* ***5****:Very confident.*

respectively. This trend is consistent with the human recognition performance shown in the next section. Interestingly, the average user confidence remains relatively stable across the different levels of syntactical complexity studied in this work.

4. Benchmark Emotion Recognition

This section reports benchmark performance on three emotion recognition tasks using the VESUS repository. Our goal is to provide a standard baseline for the community, rather than to present a new method for emotion recognition. To facilitate reproducibility, all of the training, validation, and testing splits used in this section will be released alongside the data.

4.1. Description of the Tasks

We have created three emotion recognition tasks based on our unique corpus design. Each of these tasks will be a five-way classification across the following emotional states: happy, sad, angry, fearful and neutral. As a sanity check, our first task is longitudinal speaker analysis. Namely, we train, validate, and test on a per-speaker basis to evaluate the consistency in emotional portrayals. Our second task focuses on syntactical complexity. In this case, our training and validation data will include just single words and multi-word phrases. We then test the emotion recognition performance on complete sentences. This experiment provides valuable insight about what emotional cues can be gleaned and generalized from short utterances. Third, we conduct a standard leave-one-speaker-out experiment to quantify the generalization power across speakers.

Finally, our crowd-sourced ratings provide a natural mechanism to moderate the difficulty of each recognition task. Specifically, we run the analyses using (1) all VESUS utterances, (2) only the utterances with $\geq$50% agreement across raters, and (3) only the utterances with $\geq$70% agreement across raters.

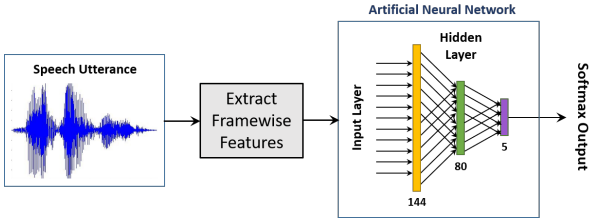

Figure 3: *Emotion classification pipeline used in this work.*

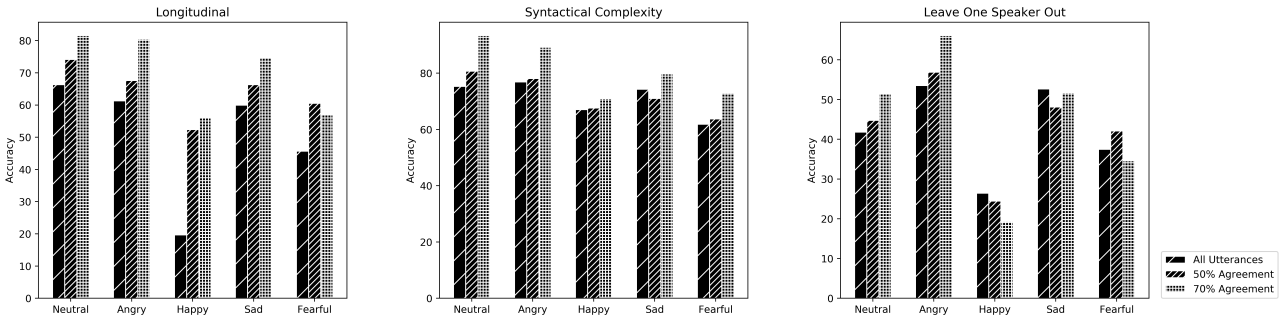

Figure 4: *Classification accuracy for each emotion across the three tasks: longitudinal speaker analysis, generalizing syntactical complexity, and leave one speaker out. Bars correspond to using all utterances, 50% rater agreement, and 70% rater agreement.*

4.2. Feature Extraction and ANN Classification

Fig. 3 outlines our simple emotion recognition pipeline. Here, we combine model-based feature selection with an artificial neural network (ANN) to segregate the emotional classes.

Our feature extraction step combines the insights from prior recognition studies [7, 21]. We segment each utterance into 400 ms frames with 50% overlap. We then extract the following frame-wise statistics as our input feature vector to the ANN:

1) **Pitch:** Fundamental frequency is computed over 50 ms sliding windows. We extract the following statistics from the F_0 signal and its first difference: mean, variance, maximum, minimum, range, kurtosis and skew. *14 features per frame*

2) **MFCC:** The first 13 MFCCs are computed over 25 ms sliding windows. The following statistics are extracted from each MFCC contour and its first difference: mean, maximum, minimum and variance. *104 features per frame*

3) **Zero-Crossing (ZC):** The ZC rate is computed over 100 ms sliding windows. We extract mean, variance, maximum, minimum (difference only), range, kurtosis and skew from this vector and its first difference. *13 features per fraome*

4) **Energy:** Squared intensity is averaged across 100 ms sliding windows. Once again, we compute the mean, variance, maximum, range, kurtosis and skew for the energy signal and its first difference. *13 features per frame*

The 144 frame-wise features are input to a fully-connected two stage ANN. The hidden layer contains 80 nodes, and the five-dimensional output layer encodes the probability of each emotional class, which we obtain via a softmax objective function. We implement the ANN using PyTorch with the Adam Optimizer for stochastic gradient descent [22]. Batch normalization is performed (batch size 32) along with 20% dropout after each layer to increase generalizability of the model. The final emotional assignment for each utterance is obtained by simple majority vote across the frame-wise probabilities.

4.3. Experimental Results

Fig. 4 illustrates the emotion recognition performance of our ANN across each of the different tasks. Notice that there is a consistent drop in accuracy for happiness, which is most pronounced in the leave-one-speaker-out experiment. Empirically, many of the happy utterances are being confused with anger due to the high arousal in both emotions. This result suggests that human portrayal of happiness is also quite variable. While not as pronounced, we also notice a dip in performance for fear, which tends to be misclassified as sadness. Both these observations are consistent with other emotional speech databases [17].

Table 4: *Emotion recognition performance for each task and level of difficulty. All results are based on the testing fold.*

	Problem Difficulty		
	All Utterances	50% Agreement	70% Agreement
Longitudinal	73.09	75.09	86.14
Complexity	56.63	67.57	75.30
Leave One Out	44.12	45.31	50.19

Another interesting observation is the generalization across syntactical complexity. In fact, Fig. 4 suggests that single words and two-word phrases contain sufficient emotional cues to learn a robust speaker-based representation. This result can be folded into real-world emotion recognition platforms, where we may only have access to fragmented speech utterances.

Finally, we note the performance gain with greater consistency in the crowd sourced annotations. Intuitively, the emotions are more pronounced in these utterances, which in turn improves machine classification. At the same time, emotional presentations are known to be variable across speakers and settings. As such, researchers can use the annotations to move from a "curated" evaluation to a more generalizable model.

5. Conclusions

We have described the design and acquisition of the new VESUS repository for emotional speech. VESUS fills a notable gap in existing resources by providing multiple examples of the same phrase being spoken with different emotional inflections. We also build syntactical complexity from single words to complete sentences, and we use crowd sourcing to quantify the perceptual variability of each utterance. These considerations allow us to define three emotion recognition tasks with varying levels of difficulty. We have provided benchmark performance on these tasks and will release the training, testing, and validation splits for further refinement by the community. Besides emotion recognition, VESUS also serves as a unique parallel database for emotional speech. Our ongoing work suggests that VESUS can be used for emotion morphing, with the ultimate goal of expressive speech synthesis [23].

6. References

[1] J. Hansen and D. Cairns, "Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments," *Speech Comm*, vol. 16, pp. 391–422, 1995.

[2] E. Hill and et al., "Long term suboxone emotional reactivity as measured by automatic detection in speech," *PLOS One*, vol. 8, pp. 1–14, 2013.

[3] P. Gupta and N. Rajput, "Two-stream emotion recognition for call center monitoring," in *Interspeech*, 2007, pp. 2241–2244.

[4] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio Speech Language Process*, vol. 17, pp. 582–596, 2009.

[5] R. Huang and C. Ma, "Toward a speaker-independent real-time affect detection system," in *International Conference on Pattern Recognition*, 2006, p. 4.

[6] E. Kim and et al., "Speech emotion recognition using eigen-fft in clean and noisy environments," in *IEEE Conference on Robot and Human Interactive Communication*, 2007, pp. 689–694.

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

[8] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *IEEE Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3687–3691.

[9] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[10] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, 2011.

[11] R. Barra-Chicote and et al., "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speechs," *Speech Communications*, vol. 52, pp. 394–404, 2010.

[12] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[13] C. Busso and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[14] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech," in *Interspeech*, 2006, pp. 805–808.

[15] S. Livingstone and F. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS ONE*, vol. 13, p. e0196391.

[16] P. Jackson and S. Haq, "SAVEE Database," http://kahlan.eps.surrey.ac.uk/savee/, Tech. Rep.

[17] C. Busso and et al., "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[18] O. Gozde and P. Daniele, "Evaluating the impact of syntax and semantics on emotion recognition from text," in *International Conference on Computational Linguistics and Intelligent Text Processing*, 2013, pp. 161–173.

[19] Education First, "English Vocabulary Lists," http://www.ef.edu/english-resources/english-vocabulary/.

[20] S. Yi, M. Steyvers, M. Lee, and M. Dry, "The wisdom of the crowd in combinatorial problems," *Cognitive Science*, vol. 36, no. 3, pp. 452–470.

[21] C. Lee and et al., "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, pp. 1162–1171, 2011.

[22] D. Kingma and J. Ba, "Adam: a method of stochastic optimization," *arXiv*, pp. 1–15, 2014.

[23] R. Shankar, J. Sager, and A. Venkataraman, "A multi-speaker emotion morphing model using deep residual networks with maximum likelihood objective," in *Submitted to Interspeech*, 2019.