DETECTING EMOTION PRIMITIVES FROM SPEECH AND THEIR USE IN DISCERNING CATEGORICAL EMOTIONS

Vasudha Kowtha, Vikramjit Mitra, Chris Bartels, Erik Marchi, Sue Booker, William Caruso, Sachin Kajarekar, Devang Naik

Apple Inc., Cupertino, CA, USA

ABSTRACT

Emotion plays an essential role in human-to-human communication, enabling us to convey feelings such as happiness, frustration, and sincerity. While modern speech technologies rely heavily on speech recognition and natural language understanding for speech content understanding, the investigation of vocal expression is increasingly gaining attention. Key considerations for building robust emotion models include characterizing and improving the extent to which a model, given its training data distribution, is able to generalize to unseen data conditions. This work investigated a long-shot-term memory (LSTM) network and a time convolution - LSTM (TC-LSTM) to detect primitive emotion attributes such as valence, arousal, and dominance, from speech. It was observed that training with multiple datasets and using robust features improved the concordance correlation coefficient (CCC) for valence, by 30% with respect to the baseline system. Additionally, this work investigated how emotion primitives can be used to detect categorical emotions such as happiness, disgust, contempt, anger, and surprise from neutral speech, and results indicated that arousal, followed by dominance was a better detector of such emotions.

Index Terms— Vocal expression, intent, paralinguistic features, long-short-term memory networks, emotion

1. INTRODUCTION

Detection of intent from a query is the principal task performed by the majority of voice operated assistants. Intent is primarily detected from words which are recognized from speech, via an automated speech recognition system. Often, voice operated assistants need to disambiguate between requests that vary in terms of vocal expression, such as an urgent query versus a casual observation. In these cases, the text recognized from the speech signal may not contain sufficient information to accurately infer the user's intent. While the traditional technological solution to this problem is to perform the disambiguation with the aid of a follow-up query, human beings are notably adept at interpreting spoken intent directly, by attending to vocal expression. Thus, this study investigates the detection of vocal expression in the form of emotion primitives through machine learning approaches.

Detection of vocal expressions in the form of emotion has received much attention in speech technology research within the past several years, where studies have focused on devising robust and relevant acoustic features [1, 2], modeling techniques [3, 4], multi-task learning [5, 6], multi-modal fusion [7, 8]. Early studies on speech-based emotion detection have used acted or elicited emotion datasets [9, 10] (where actors were recorded while speaking with specified emotions). Observations from acted-emotion studies revealed that models trained with acted emotions may not generalize well to spontaneous subtle emotions [11] as a consequence, datasets containing spontaneous emotions were collected, such as the MSP Podcast dataset [12]. The downside of collecting spontaneous speech emotion datasets is that they typically lack ground-truth labels, and thus require manual annotation, that suffer from varying degrees of grader agreement. Labels in emotion datasets contain either categorical emotions (such as happy, sad, neutral etc.) and/or primitive emotions (such as valence, arousal, dominance etc.). While categorical emotions are easy to interpret, they are difficult to annotate, as they often lead to annotator disagreements, skewed datasets and suffer from ambiguity in defining the lexicon for emotion categories [13]. On the other hand, primitive emotions, which are defined by the valence-arousal-dominance scale, are easier to annotate but harder to interpret, and typically generate results that are more easily comparable. Moreover, primitive emotions can be coarsely aggregated, to recover categorical emotions [14].

In this work we used spontaneous speech corpus labelled with primitive emotions and investigated the following:
(1) Role of annotator agreement on model performance.
(2) Role of acoustic features on model robustness.
(3) Whether such models are useful in detecting vocal expression in the form of happiness, disgust, contempt, anger, and surprise.

Through our work we demonstrate that:
(a) Use of low-dimensional frame-level features (such as filterbank energies) can demonstrate performance comparable to complex feature sets investigated in the literature.
(b) Presence of additional data resources can help to improve an emotion detection model's performance and generalization capacity.

(c) Simple score level fusion of multiple complimentary systems can improve the overall performance

The outline of the paper is as follows: section (2) will present the datasets used in our study, (3) will introduce the acoustic features investigated in this work, (4) will detail the acoustic model and its parameters, in (5) we will present the results, followed by conclusion in (6).

2. DATA

We use a slightly expanded version of the data used in our earlier study [13] which contains 120 hours speech material spoken in US English. The data had no speaker level information. The duration of each utterance varied between 2 to 6 seconds. The data contained perceptually assigned valence, arousal and dominance scores. For more information on this dataset, please refer to [13]. Additionally, this study also uses the MSP-Podcast data [12] that contains speech spoken by English speakers collected from online audio shows, covering topics such as politics, sports, entertainment, etc. The speech segments in this dataset contain single speaker utterances with duration between 2.75 and 11 seconds. Overall the MSP-Podcast (ver-3) data contained a little over 50 hours of speech. The data came with speaker and gender labels, which were not used in this study. There were altogether 588 speaker labels, where 50 speakers were present in the test set. The remaining data was used to train the model, where 90% of the data was used as the training set and the remaining 10% as the cross-validation set. We will denote the MSP-Podcast training data as MSP-train. To make our results comparable to the literature, we will report results on MSP-Podcast (version 2) eval set ($\approx$ 12 hours of speech), which we will denote as MSP-eval.

3. ACOUSTIC FEATURES

We investigated multiple acoustic features to parameterize speech. The baseline feature is the 40-dimensional mel-filterbank energy (MFB) features, appended by pitch, pitch-delta and voicing features, which we denote as $MFB + F0$ feature. We explored 40-D gammatone filter-bank energies (GFB) and speech modulation energies (extracted through the amplitude modulation (MOD) feature extraction setup as specified in [15]), both were appended with 3-D pitch ($F0$) and voicing features and we denote them as $GFB + F0$ and $MOD + F0$ features. We have used articulatory features in the form of vocal-tract constriction variables (TV) [16]. In our earlier work [13] we have shown that TVs can assist in detecting valence from the speech signal. The TVs define degree and location of constriction actions within the human vocal tract and have eight dimensions [16, 17]. Similar to [13] we have used an LSTM-based speech-inversion system which takes in spliced (window of 5 frames on both sides of the current frame) $MFB + F0$ features as input and maps that to the 8 TV trajectories.

4. ACOUSTIC MODELING

We have used single-layer LSTM networks consisting of 128 neurons in the recurrent and the embedding layers, to train the baseline primitive emotion detection (regression) model. The input to the model was low-level features described in section 4, which were analyzed with a window size of 25 ms and a frame rate of 10 ms, and the output was 3-D primitive emotions: valence, arousal and dominance. The model was tuned using a held-out dev set, and based upon that the number of neurons in each layer and the cost function (concordance-correlation-coefficient, CCC_{cost} as shown in (1)) was selected. The CCC_{cost} is a combination ($\alpha = 1/3$ and $\beta = 1/3$) of CCC's obtained from each of the valence, dominance and arousal dimensions. CCC for each dimension is defined by (2), where where μ_x and μ_y are the means, σ_x^2 and σ_y^2 are the corresponding variances for the estimated and ground truth variables and ρ is the correlation coefficient between those two variables. The models were trained with a mini-batch size of 512, using Adam optimizer, with a learning rate of 0.001. For all the model training steps, early stopping was allowed based on cross-validation error.

$$CCC_{cost} := \alpha CCC_{val} + \beta CCC_{aro} + (1 - \alpha - \beta) CCC_{dom} \quad (1)$$

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (2)$$

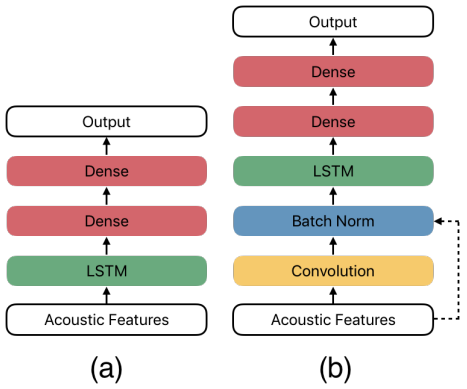

Fig. 1. Architecture of (a) LSTM network and (b) Time convolution LSTM network (TC-LSTM)

Additionally, we investigated a time-convolutional layer before the LSTM layer (TC-LSTM, shown in Figure 1), where the number of convolutional layers (having filter size = 3) was same as the number of input feature dimensions. The network also had a skip connection, where acoustic features were also fed directly to the LSTM layer in addition to the convolutional layer outputs. The LSTM layer had 128 and 312 neurons in the recurrent and embedding layers, respectively.

5. RESULTS AND ANALYSIS

We trained three LSTM models with (a) internal-train, (b) MSP-train and (c) internal+MSP train data, which were evaluated with internal-eval and MSP-eval data respectively. These models were trained with $MFB+F0$ features and the results are shown in table 1.

Table 1. Primitive emotion CCC from models trained with baseline MFB+F0 features

	Internal-eval			MSP-eval		
	val	aro	dom	val	aro	dom
Int-train	0.56	0.69	**0.58**	0.13	0.53	0.40
MSP-train	0.14	0.55	0.47	0.24	0.73	**0.68**
Both-tr	**0.57**	**0.70**	**0.58**	**0.26**	**0.75**	**0.68**

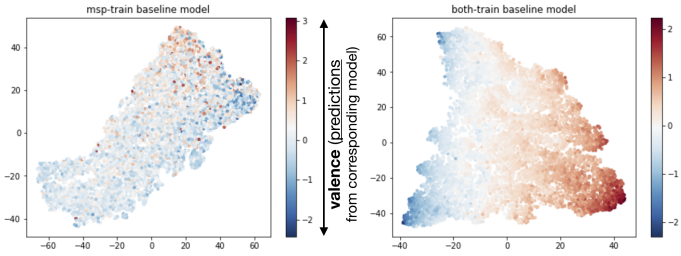

Fig. 2. t-SNE plot of embeddings from MSP-train and Both-tr models, colored w.r.t valence score

Table 1 shows the impact of data mismatch between training and evaluation data. Model trained with the internal-train data did better on the internal-eval set, while the model trained with MSP-train, performed better for MSP-eval set, and both failing to demonstrate similar performance on the mismatched eval set (MSP-eval and internal-eval sets, respectively). This provides some realistic interpretation of how much performance gets impacted due to domain mismatch, which is found to be $\approx 40\%$ reduction in CCC. The last row in Table 1 shows that a model trained with internal-train and MSP-train data (followed by fine-tuning with each of those training sets) demonstrated a much better performance for both the eval sets. Given, that the Both-train model gave the best baseline CCC performance for both the eval sets, we will be using that model as our baseline in the rest of this paper. Figure-2 shows the t-SNE plot of the embeddings obtained from MSP-train and Both-train models for different values of valence, where we can observe that the latter model has a better separation of high-valence data points from low-valence ones.

5.1. Role of Annotator Consensus

A careful analysis of the evaluation set indicated that annotation consensus played an important role on evaluation performance. The MSP-podcast data came with annotator decisions on the primary categorical emotion, where the number of annotators grading a specific utterance varied from five to sixteen. The primary categorical emotion consensus on the evaluation set is determined by the number of annotators who selected category that received the majority vote divided by the total number of annotators who voted for that utterance [18]. We denote an utterance's consensus ($p(C)$) by the probability of an annotator selecting the primary categorical emotion that received the majority vote [18]. Table 2 presents the results from the MSP-eval set as obtained from the baseline model, when grouped by annotator consensus.

Table 2 presents some interesting observations on the role of annotator consensus on performance evaluation. While valence CCC always increased with increase in consensus (last three rows in table 2), arousal and dominance remained relatively stable. This may indicate that valence may be relatively difficult to annotate and can be correlated with the consensus emotion decision, compared to arousal and dominance. Interestingly, at extremely low consensus all of the three primitive emotions demonstrated extreme deterioration in performance, indicating low-consensus data may not be reliable in assessing a model's performance. From these observations we can claim that:

(1) Primitive emotion models are sensitive to annotator consensus, where valence was found to be more sensitive to it compared to arousal and dominance.

(2) Data with less than 25% consensus may not be suitable for assessing the goodness of a model, and it may be useful to not consider such data points.

Table 2. Primitive emotion CCC from multi-condition trained baseline model trained when MSP-eval set is grouped by annotator consensus

Annotator Consensus	MSP-eval (CCC)		
	val	aro	dom
$p(C) \leq 0.25$	0.00	0.59	0.57
$p(C) \leq 0.40$	0.12	0.73	0.67
$p(C) \geq 0.50$	0.32	0.77	**0.70**
$p(C) \geq 0.60$	0.33	0.77	0.69
$p(C) \geq 0.75$	0.40	**0.78**	**0.70**
$p(C) \geq 0.90$	**0.46**	**0.78**	0.69

Table 3. Primitive emotion CCC from LSTM models trained with different acoustic features

	Internal-eval			MSP-eval		
	val	aro	dom	val	aro	dom
$MFB+F0$	**0.57**	**0.70**	**0.58**	0.26	**0.75**	**0.68**
$GFB+F0$	0.56	0.68	0.57	0.29	0.72	0.63
$MOD+F0$	0.56	0.69	0.56	**0.30**	0.72	0.63

5.2. Robustness

To investigate if model performance can be improved beyond the baseline, we trained several models using $GFB+F0$ and $MOD+F0$ features. The results from that study is shown

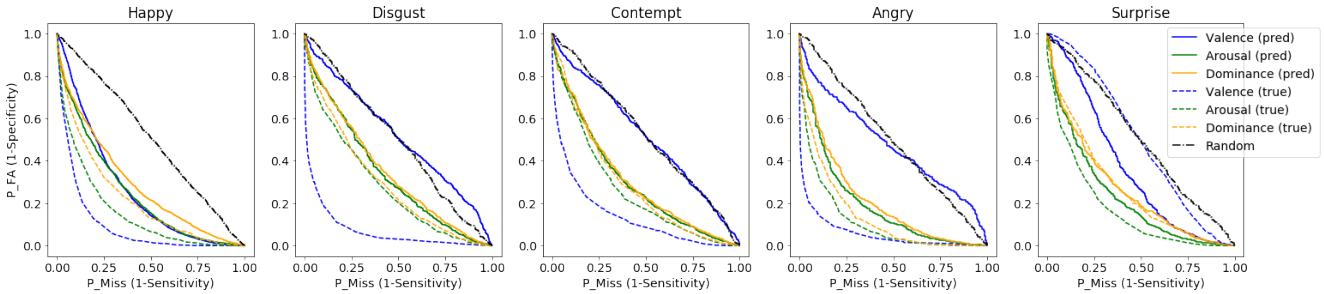

Fig. 3. ROC curves for detecting categorical emotions: happy, disgust, contempt, angry and surprise from neutral.

Table 4. Primitive emotion CCC from TC-LSTM models trained with different acoustic features

	Internal-eval			MSP-eval		
	val	aro	dom	val	aro	dom
$MFB + F0$	0.56	0.69	0.57	0.27	**0.75**	**0.69**
$GFB + F0$	0.58	**0.70**	**0.58**	0.32	0.74	0.65
$MOD + F0$	0.57	**0.70**	**0.58**	0.31	0.74	0.67
$MOD + TV + F0$	**0.59**	**0.70**	**0.58**	**0.33**	0.74	0.68

Table 5. Primitive emotion CCC from TC-LSTM models and the relevant state-of-the-art

	MSP-eval			
	val	aro	dom	Params
CNN MFB [3]	0.25	0.74	0.66	800K
Best in [3]*	0.30	**0.77**	**0.70**	>1M
TC-LSTM $MODTV + F0$	0.33	0.74	0.68	100K
TC-LSTM $MOD + TV + F0$ + TC-LSTM $GFB + F0$	**0.34**	**0.77**	0.69	200K

below in Table 3. Table 3 shows that the $GFB + F0$ and $MOD + F0$ demonstrated comparable performance with respect to the baseline $MFB + F0$ features, but with improved (statistically significant) CCC for valence on the MSP-eval sets. Table 4 shows the results obtained from the TC-LSTM acoustic model, where it can be seen that the TC-LSTM overall performed better than the LSTM model. We observed that the best result from our TC-LSTM model is better than a comparable MFB-CNN model (multi-task learning) that was evaluated on the same MSP_eval set (shown in Table 5), and was very close to the best performing (* in Table 5) system that used more than 6K feature.

5.3. Application: Detection of Categorical Emotions from their primitives

To investigate how primitive emotion decisions generalize to categorical emotions, we investigated the task of detecting happiness, disgust, contempt, anger and surprise versus neutral, given the valence, arousal and dominance scores from the $MFB + F0$ LSTM model. Figure 3 shows the ROC curves for detecting the respective categorical emotions from neutral, given the predictions (*pred*) from the $MFB + F0$ LSTM model and the *true* label. Table 6 presents the area under the curve (AUC) for each of these cases and shows that while arousal (*pred*) is a strong indicator for all the categories, valence is useful for detecting happiness, dominance for disgust, contempt and anger. ROC in Figure 3 shows that the (*true*) valence is the best indicator for detecting happiness, disgust, contempt and anger however (*pred*) valence shows much worse performance compared to it; which motivates the necessity to improve the detection of valence from speech.

Table 6. Primitive Emotion AUC from TC-LSTM models trained with different acoustic features

	AUC				
	Happy	Disgust	Contempt	Anger	Surprise
Val	0.36	0.48	0.49	0.52	0.36
Aro	**0.34**	**0.33**	**0.30**	**0.19**	**0.22**
Dom	0.39	0.35	**0.30**	0.21	0.26

6. CONCLUSION

We investigated a TC-LSTM model to detect primitive emotions from speech, and demonstrated that it performed significantly better in detecting valence compared to the state-of-the-art reported in the literature for a publicly available dataset. We observed that $MOD+TV+F0$ features offered the best performance for detecting valence, while $MFB + F0$ features performed better for detecting arousal and dominance. We demonstrated that frame-level filterbank energy features can generate comparable performance to that of large dimensional features typically used in the literature. We also observed that simple score level fusion can improve overall emotion detection performance. Finally, we observed that the model generated primitive emotion scores are useful to detect categorical emotions.

7. ACKNOWLEDGEMENT

The authors would like to thank Russ Webb and Panayiotis Georgiou for their valuable comments and suggestions.

8. REFERENCES

[1] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, MM '13, pp. 835–838, ACM.

[2] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Phuong Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing"," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 4 2016, Open access.

[3] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM transactions on audio, speech, and language processing*, pp. 1–13, May 2019.

[4] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, pp. 2203–2213, 2014.

[5] Reza Lotfian and Carlos Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *INTERSPEECH*, 2018.

[6] Rui Xia and Yang P. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, vol. 8, pp. 3–14, 2017.

[7] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, New York, NY, USA, 2004, ICMI '04, pp. 205–211, ACM.

[8] Ashish Kapoor and Rosalind W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, New York, NY, USA, 2005, MULTIMEDIA '05, pp. 677–682, ACM.

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[10] Philip Jackson and Sana ul haq, "Surrey audio-visual expressed emotion (SAVEE) database," 2014.

[11] Ellen Douglas-Cowie, Laurence Devillers, Jean-Claude Martin, Roddy Cowie, Suzie Savvidou, Sarkis Abrilian, and Cate Cox, "Multimodal databases of everyday emotion: facing up to complexity," *INTERSPEECH*, pp. 813–816, 2005.

[12] Soroosh Mariooryad, R. Lotfian, and Carlos Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 238–242, 01 2014.

[13] Vikramjit Mitra, Sue Booker, Erik Marchi, David Farrar, Ute Peitz, Bridget Cheng, Ermine Teves, Anuj Mehta, and Devang Naik, "Leveraging acoustic cues and paralinguistic embeddings to detect expression from voice," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 06 2019.

[14] Michael Grimm, Emily Mower, Kristian Kroschel, and Shrikanth Narayanan, "Combining categorical and primitives-based emotion recognition," in *2006 14th European Signal Processing Conference*. IEEE, 2006, pp. 1–5.

[15] Vikramjit Mitra, Horacio Franco, Martin Graciarena, and Arindam Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," 03 2012, pp. 4117–4120.

[16] Vikramjit Mitra, Ganesh Sivaraman, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman, and Mark Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Communication*, vol. 89, pp. 103–112, 2017.

[17] Vikramjit Mitra, Hosung Nam, Carol Y Espy-Wilson, Elliot Saltzman, and Louis Goldstein, "Retrieving tract variables from acoustics: a comparison of different machine learning strategies," *IEEE journal of selected topics in signal processing*, vol. 4, no. 6, pp. 1027–1045, 2010.

[18] Wei Tang and Matthew Lease, "Semi-supervised consensus labeling for crowdsourcing," in *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR), 2011*, 2011.