Acoustic Feature Analysis in Speech Emotion Primitives Estimation

Dongrui Wu[1,2], *Thomas D. Parsons*[1], *Shrikanth S. Narayanan*[2]

[1]University of Southern California, Institute for Creative Technologies, Marina del Rey, CA, USA
[2]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
dongruiw@usc.edu, tparsons@ict.usc.edu, shri@sipi.usc.edu

Abstract

We recently proposed a family of robust linear and nonlinear estimation techniques for recognizing the three emotion primitives–valence, activation, and dominance–from speech. These were based on both local and global speech duration, energy, MFCC and pitch features. This paper aims to study the relative importance of these four categories of acoustic features in this emotion estimation context. Three measures are considered: the number of features from each category when all features are used in selection, the mean absolute error (MAE) when each category is used separately, and the MAE when a category is excluded from feature selection. We find that the relative importance is in the order of *MFCC* $>$ *Energy* $\approx$ *Pitch* $>$ *Duration*. Additionally, estimator fusion almost always improves performance, and locally weighted fusion always outperforms average fusion regardless of the number of features used.

Index Terms: Emotion estimation, 3D emotion space, speech analysis, estimator fusion, support vector regression, robust regression, locally linear reconstruction, locally weighted fusion

1. Introduction

Emotions may be recognized from many different information sources, e.g., speech [1–3], facial expressions [4,5], physiological signals [6,7], or their multimodal combination [8,9]. In this paper we focus on emotion recognition from speech signals.

A majority of research on speech emotion recognition classifies emotions into a small number of categories [1, 10, 11]. Emotion psychology research [12–14] has shown that emotions can also be represented as points in a multi-dimensional space, i.e., emotions can be quantified as continuous numbers instead of categorical values. One of the most frequently used emotion spaces consists of three primitives [14, 15] of *Valence*, *Activation*, and *Dominance*. This 3D representation is easier to implement than categories because, though computing with words [16, 17] is possible, computers are better at dealing with numbers for deriving inferences and decision making. This 3D representation of emotions is used in this paper.

We [3] have recently introduced three elementary models and two fusion approaches for estimating speech emotions in 3D space using the VAM corpus [18]. Four categories of acoustic features widely adopted in the literature (duration, energy, MFCC, and pitch) were used. The three elementary models had comparable performance with the state-of-the-art results [2,19], and the two fusion models outperformed them. This paper aims to provide more insights on which features are more important for speech emotion estimation in the 3D space using the newly proposed estimators. Particularly, it studies the relative importance of the four categories of acoustic features using three different measures.

Feature importance is a key topic in speech emotion recognition. It has been investigated in a number of papers [11, 20–24]. However, all of them study the importance of acoustic features in classifying emotions into categories. In contrast, this paper investigates feature importance in estimating the continuous values of emotions in 3D space, a topic that has not been widely addressed. Our novel contributions are as follows:

1. We introduce three measures to evaluate the relative importance of the four categories of acoustic features in emotion estimation.
2. We use three elementary estimators in each of the three measures, and also two fusion models in the latter two measures.

The rest of this paper is organized as follows: Section 2 briefly introduces our experiment setup. Section 3 presents our experimental results. Finally, Section 4 draws conclusions and proposes some future works.

2. Experiment Setup

The VAM corpus [18] was used in this paper. It contains spontaneous speech with authentic emotions recorded from guests in a German TV talk-show *Vera am Mittag* (*Vera at Noon* in English). There are 947 emotional utterances from 47 speakers (11m/36f). Each utterance was evaluated by 6-17 human listeners in the 3D space of valence, activation and dominance. Then, their evaluations in each dimension were aggregated to obtain a number in [-1, 1]. So, the emotion of each utterance is represented by a 3D vector, which serves as our reference.

46 acoustic features, which are listed in the second column of Table 1, were extracted. They are the same as those used in [2,3] on the same corpus and cover four major categories:

- *Duration features* (5): mean and standard deviation of the duration of voiced and unvoiced segments, ratio between the duration of unvoiced and voiced segments.
- *Energy features* (6): energy mean, standard deviation, maximum, 25% and 75% quantiles, and the inter-quantile distance.
- *MFCC features* (26): mean and standard deviation of 13 Mel frequency cepstral coefficients (MFCC).
- *Pitch features* (9): f0 mean, standard deviation, median, minimum, maximum, range, 25% and 75% quantiles, and the inter-quantile distance.

Three elementary estimators – robust regression (RR), support vector regression (SVR), and locally linear reconstruction

This research is partially supported by U.S. Army Research Laboratorys Translational Neurosciences Branch (Aberdeen), and the National Science Foundation.

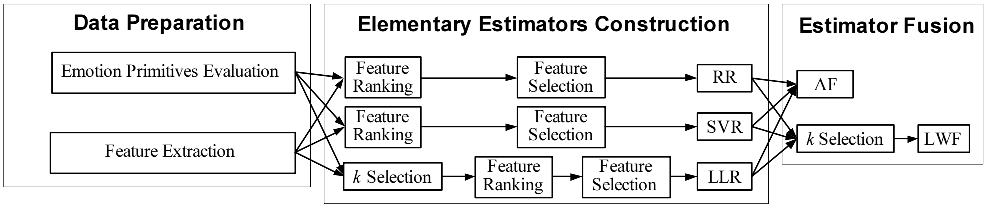


Figure 1: A flowchart for speech emotion primitives estimation.

(LLR), and two estimator fusion approaches – average fusion (AF) and locally weighted fusion (LWF), were constructed, as shown in Fig. 1. They are exactly the same as those reported in our previous study [3]. The three elementary estimators were chosen because they are diverse and complementary, in the sense that they cover both local model (LLR) and global models (RR and SVR), and both linear models (RR and LLR) and nonlinear model (SVR). So, we expect better performance when they are fused properly. AF takes the average of the three elementary estimators. LWF computes the weighted average of RR, SVR and LLR, where the weights are adaptive and dependent on the local performance of the elementary estimators.

For feature selection, we first ranked the features using the iterative sequential backward selection method [25], and then performed cross-validation according to the rank of the features to select the best subset. We started with all 46 features and gradually removed the worst feature until the 10-fold cross-validation performance stopped improving.

3. Experimental Results

The mean absolute error (MAE) between the estimates, $\hat{y}_n$, and the human evaluations, y_n, i.e., MAE $= \frac{1}{947}\sum_{n=1}^{947}|\hat{y}_n - y_n|$, was used in performance evaluation. It has also been used in [2,3,19]. Three measures based on MAE were employed in this paper:

1. *Measure1*: The number of features from each category when all four categories are used in feature selection.
2. *Measure2*: The MAE when each category of features are used separately in feature selection.
3. *Measure3*: The MAE when a category is excluded from feature selection (the remaining three categories are used together).

Experimental results on the three measures are presented next.

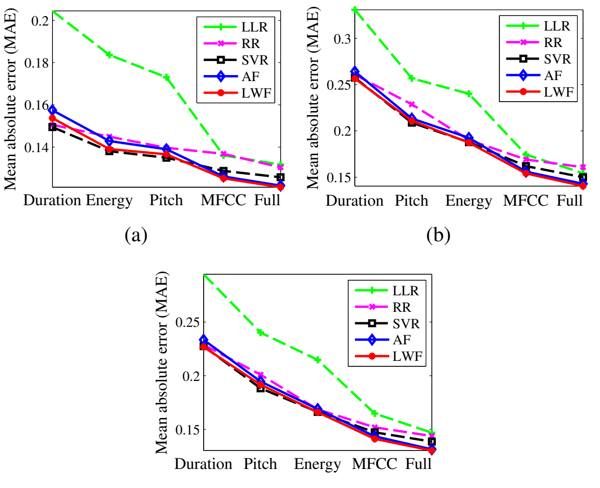
3.1. Experimental Results for Measure1

For Measure1, we used the number of features as an importance indicator and assume that the more features are selected from a certain category, the more important that category is. The features used by RR, SVR and LLR for estimating the three emotion primitives are shown in Table 1. Observe that:

1. Generally all four categories of features contributed in emotion primitives estimation. This is consistent with Batliner et al.'s [21] observation when these categories were used in emotion classification.
2. The optimal feature subsets for the three elementary estimators were different for each emotion primitive.
3. For RR, the importance of the four categories was in the order of *MFCC* > *Pitch* > *Energy* > *Duration*.

Table 1: The features used by RR, SVR and LLR for estimating the three emotion primitives. The rank of a feature in each category is determined by the number of times it appeared in the nine cases (shown in the last column); so, generally a higher rank indicates higher importance.

		RR			SVR			LLR			
Category	Feature	V	A	D	V	A	D	V	A	D	#
Duration Features	pauseDurationStd		√	√		√		√	√	√	6
	speechDurationMean					√		√	√		3
	pauseDurationMean							√	√	√	3
	pause2SpeechRatio									√	1
	speechDurationStd										0
Energy Features	intensityMax	√	√	√	√			√	√	√	7
	intensityQ75		√		√	√		√	√		5
	intensityStd				√	√	√			√	4
	intensityQRange			√		√	√			√	4
	intensityMean							√	√	√	3
	intensityQ25						√	√		√	3
Pitch Features	f0Min	√	√	√		√		√		√	6
	f0Q75			√		√	√	√	√	√	6
	f0Median	√			√	√		√		√	5
	f0Q25				√	√		√	√	√	5
	f0Mean		√			√		√	√		4
	f0Max		√	√	√						3
	f0QRange				√	√				√	3
	f0Range		√	√							2
	f0Std					√					1
MFCC Features	mfccMean1	√	√	√	√	√	√	√	√	√	9
	mfccMean2		√	√	√	√	√	√	√	√	8
	mfccStd3		√	√	√	√	√	√	√	√	8
	mfccMean3	√	√	√		√	√	√	√		7
	mfccMean7	√	√		√	√		√	√	√	7
	mfccMean9	√	√	√		√		√	√	√	7
	mfccStd4	√			√	√	√	√	√	√	7
	mfccMean4		√		√	√		√	√	√	6
	mfccMean6	√	√	√				√	√	√	6
	mfccMean8		√			√	√	√	√	√	6
	mfccMean12		√			√	√	√	√	√	6
	mfccStd2	√			√	√		√	√	√	6
	mfccStd5		√			√	√	√	√	√	6
	mfccStd10					√	√	√	√	√	5
	mfccMean5		√				√	√	√		4
	mfccMean10			√	√			√		√	4
	mfccMean11				√			√	√	√	4
	mfccStd1				√		√	√		√	4
	mfccStd13		√		√			√	√		4
	mfccMean13	√					√	√			3
	mfccStd7				√			√		√	3
	mfccStd8					√		√	√		3
	mfccStd6				√			√			2
	mfccStd12					√		√			2
	mfccStd9							√			1
	mfccStd11							√			1



(a) (b)

(c)

Figure 2: MAEs of the five estimators when the four categories of features were used separately. For example, "Duration" means only the five duration features were used in feature selection. "Full" means all 46 features from the four categories were used. A larger MAE means that category of features are less important. (a) Valence; (b) Activation; and, (c) Dominance.

4. For SVR, the importance of the four categories was in the order of *MFCC* > *Pitch* > *Energy* > *Duration* for Valence and Activation, and *MFCC* > *Energy* > *Pitch* > *Duration* for Dominance.
5. For LLR, the importance of the four categories was in the order of *MFCC* > *Energy* ≈ *Pitch* > *Duration*.

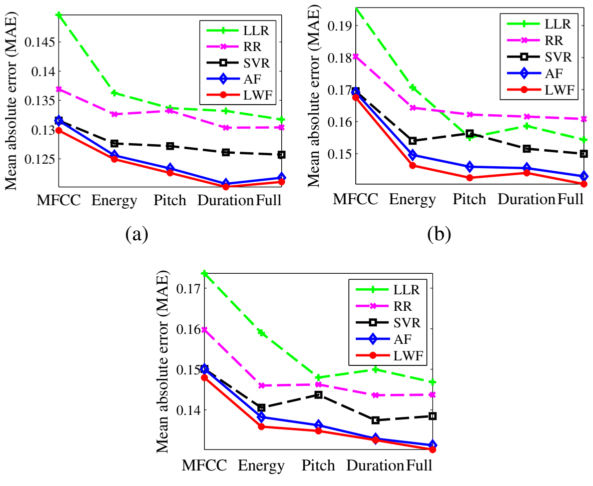
In summary, the importance of the four categories was in the order of *MFCC* > *Energy* ≈ *Pitch* > *Duration*.

3.2. Experimental Results for Measure2

For Measure2, each category of features was used separately in feature selection. So, by comparing the MAEs of the models obtained from different categories, the importance of the four categories is directly evaluated. A comparison of the performances is shown in Fig. 2. Note that here a smaller MAE means a better performance and hence higher importance. Observe that:

1. For Valence, the importance of the four categories was in the order of *MFCC* > *Pitch* > *Energy* > *Duration*.
2. For Activation and Dominance, the importance of the four categories was in the order of *MFCC* > *Energy* > *Pitch* > *Duration*.
3. SVR always had the smallest MAE among the three elementary estimators, and LLR often had the largest.
4. When the number of features was small (i.e., when duration, energy, and pitch were used separately), LLR had much worse performance than RR and SVR, and fusion did not necessarily reduce the MAE.
5. LWF always outperformed AF, regardless of the number of features used.

In summary, the importance of the four categories was in the order of *MFCC* > *Energy* ≈ *Pitch* > *Duration*, which is consistent with our finding in Measure1.



(a) (b)

(c)

Figure 3: MAEs of the five estimators when one category of features were excluded. For example, "MFCC" means the 26 MFCC features were excluded in feature selection whereas all other 20 features were used. "Full" means all 46 features from the four categories were used. A larger MAE means that category of features are more important. (a) Valence; (b) Activation; and, (c) Dominance.

3.3. Experimental Results for Measure3

For Measure3, each time we excluded one category of features from feature selection and used the remaining three categories. The MAE in Measure3 is an indirect importance indicator, in contrast to Measure2, because here the MAE is an indicator of performance loss if a category of features are not used. However, Measure3 may be more meaningful than Measure2 since in practice we usually combine features from different categories, and Measure3 indicates how much new useful information a category can add to the existing feature combination.

A comparison of the performances is shown in Fig. 3. Note that for Measure3 a larger MAE means higher importance, because a larger MAE indicates that the estimation performance downgrades more when that category of features are excluded. Observe that:

1. For the three elementary estimators, MFCC was always the most important. Generally, Duration was the least important. Pitch and Energy were moderately important; however, there was no clear evidence that one was more important than the other.
2. For the two fusion models, the importance of the four categories was in the order of *MFCC* > *Energy* > *Pitch* > *Duration*.
3. Fusion models almost always outperformed the elementary models, and LWF always outperformed AF.

In summary, the importance of the four categories was in the order of *MFCC* > *Energy* ≈ *Pitch* > *Duration*, which is consistent with our findings in Measure1 and Measure2.

4. Conclusions and Future Works

In this paper, we have compared the relative importance of four categories of acoustic features in speech emotion estimation, using three elementary estimators and two fusion models. The main findings are:

1. For all three emotion primitives and all five models, MFCC features were always the most important. This finding is consistent with Vogt and Andre's observation [20] in emotion classification using spontaneous speech, and also Schuller and Rigoll's observation [24] in emotion recognition using both frame-level and supra-segmental features.
2. Often Duration features were the least important.
3. Pitch and Energy features were moderately important; however, there was no evidence that one was more important than the other.
4. In summary, the importance of the four categories was in the order of $\mathit{MFCC} > \mathit{Energy} \approx \mathit{Pitch} > \mathit{Duration}$. This pattern is consistent with Mower et al.'s findings [23], where the same 46 acoustic features were applied to two datasets in English and information gain was used in feature selection for emotion classification.
5. Generally, estimator fusion can achieve better performance than the elementary estimators. Between the two fusion approaches, LWF always outperformed AF, regardless of how many features were used.

We need to point out that only 46 acoustic features were used in this study, and the four categories had different numbers of features, i.e., MFCC, Energy, Pitch and Duration had 26, 6, 9, and 5 features, respectively. It is interesting to investigate whether or not extracting more and different features in these four categories will change their relative importance and improve the estimation performance. For example, the 137 acoustic features used by Grimm and Kroschel [19], and the 3304 acoustic features used by Batliner et al. [21], can be considered. Additionally, Busso et al. [26] showed that Mel Filter Bank (MFB) features seem more important than MFCC features in emotional speech analysis. It is interesting to augment MFB features in our study. Finally, we can extend this study to databases in different languages and study whether the pattern found in this paper is universal.

5. References

[1] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[2] M. Grimm, K. Kroschel, E. Mower, and S. S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[3] D. Wu, T. D. Parsons, E. Mower, and S. S. Narayanan, "Speech emotion estimation in 3D space," in *Proc. IEEE Int'l Conf. on Multimedia & Expo (ICME)*, Singapore, July 2010.

[4] M. Pantic and L. J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[5] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[6] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with Computers*, vol. 21, pp. 133–145, 2009.

[7] J. Kim and E. Andre, "Fusion of multichannel biosignals towards automatic emotion recognition," in *Multisensor Fusion and Integration for Intelligent Systems*, ser. Lecture Notes in Electrical Engineering, S. Lee, H. Ko, and H. Hahn, Eds. Berlin Heidelberg: Springer-Verlag, 2009, pp. 55–68.

[8] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. Int'l Conf. on Multimodal Interfaces*, State Park, PA, October 2004, pp. 205–211.

[9] A. Metallinou, S. Lee, and S. S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Proc. Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, March 2010, pp. 2462–2465.

[10] N. F. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[11] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157–183, 2003.

[12] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological Review*, vol. 110, pp. 145–172, 2003.

[13] H. Schlosberg, "Three dimensions of emotion," *Psychological Review*, vol. 61, no. 2, pp. 81–88, 1954.

[14] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, 2003.

[15] R. Kehrein, "The prosody of authentic emotions," in *Proc. Speech Prosody Conf.*, Aix-en-Provence, France, April 2002, pp. 423–426.

[16] L. A. Zadeh, "Fuzzy logic = Computing with words," *IEEE Trans. on Fuzzy Systems*, vol. 4, pp. 103–111, 1996.

[17] J. M. Mendel and D. Wu, *Perceptual Computing: Aiding People in Making Subjective Judgments*. Hoboken, NJ: Wiley-IEEE Press, 2010.

[18] M. Grimm, K. Kroschel, and S. S. Narayanan, "The Vera Am Mittag German audio-visual emotional speech database," in *Proc. Int'l Conf. on Multimedia & Expo (ICME)*, Hannover, German, June 2008, pp. 865–868.

[19] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3D emotion space concept," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech, 2007, pp. 281–300.

[20] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. IEEE Int'l Conf. on Multimedia & Expo (ICME)*, Amsterdam, The Netherlands, July 2005, pp. 474–477.

[21] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. A. L. Kessous, and N. Amir, "Whodunnit – searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech and Language*, 2010, in press.

[22] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. InterSpeech*, Antwerp, Belgium, August 2007, pp. 2253–2256.

[23] E. Mower, M. J. Mataric, and S. S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 843–855, 2009.

[24] B. Schuller and G. Rigoll, "Recognising interest in conversational speech – comparing bag of frames and supra-segmental features," in *Proc. InterSpeech*, Brighton, UK, September 2009, pp. 1999–2002.

[25] K. S. Fu, *Sequential Methods in Pattern Recognition and Machine Learning*. NY: Academic Press, 1968.

[26] C. Busso, S. Lee, and S. S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proc. InterSpeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.