ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Emotional Speech: A Spectral Analysis

Pouria Fewzee, Fakhri Karray

Centre for Patter Analysis and Machine Intelligence,
University of Waterloo, Waterloo, ON, Canada
{spfewzee, karray}@uwaterloo.ca

Abstract

Feature extraction and dimensionality reduction may be found as the most imperative parts of emotional speech recognition problem. In this work, we propose a new set of speech features based on the distribution of energy in frequency domain. To investigate the applicability of the proposed model, we have set the first international audio/visual emotion challenge (AVEC 2011) as a benchmark. As for the modeling and dimensionality reduction, we have employed the lasso. It is shown how 15 explicit spectral energy features, as suggested in this work, can lead to a more accurate model than those of the participants in the audio sub-challenge. This is while this number of features is less than ten percent of the smallest set of features in the challenge.

Index Terms: emotional speech recognition, feature extraction, dimensionality reduction

10.21437/Interspeech.2012-130

1. Introduction

Emotional speech recognition is the problem of discriminating speech samples by their emotional content. To tackle this problem, one should consider a few matters [1, 2]. One is the presentation of emotion, whether to consider dimensional or categorical emotions. Next, the choice of speech features plays an important role in a solution for the problem. This by itself involves dimensionality reduction matters. Finally, the choice of statistical model, for capturing patterns of variation between extracted features and emotional contents of speech, is of a high importance. In this work, we set feature extraction and dimensionality reduction as our main concerns and rely on the previous works, when it comes to the other matters.

A major part of the recent work has been directed towards the objective of the first international audio/visual emotion challenge and workshop in 2011 (AVEC 2011) [3]. In the context of three sub-challenges, participants were expected to predict emotional contents of speech in word level and based on four binary dimensions: activity, expectation, power, and valence. The three sub-challenges are run based on voice cues (audio), facial cues (visual), and the fusion of both (audio/visual). That is to say, to use audio and/or visual signals to predict whether each word in a conversation has a low or high extent in each dimension. In this work, we adopt the audio sub-challenge as for the framework of our experiments. This enables us to make a comparison between our approach and the most recent advances in the field. Here, we go briefly over some of the works done by the participants in the audio sub-challenge.

In order to deal with dimensionality reduction, Calix et al [4] have used a Chi-square ranking process to select a subset of features. In a work by Cen et al [5], one can see the use of the ℓ_1 norm, employed in a regression formulation, to select an optimal subset of speech features. Glodek et al [6] employ a multiple classifier system for recognizing emotional states, based on voice and facial cues. Focusing mainly on dimensionality reduction, Kim et al [7] apply maximum average recall, maximum relevance, and minimal-redundancy-maximal-relevance for feature selection. Pan and others [8] have used PCA to reduce dimensionality, along with SVM and AdaBoost for learning. Proposed by Sayedelahl et al [9] is two sets speech features. One is co-occurrence matrix for the extraction of meta features from speech and the other one is a sense of energy distribution in the frequency domain. In a work by Sun and Moore [10], the applicability of two new sets of features, namely glottal waveform parameters and Teager's energy operators, has been investigated.

Looking at the most recent works, one may notice that a major focus is on the recognition of an optimal set of speech features. When we say optimal, we mean to address two characteristics of a set of features: to be least in number, but most in conveyance. In this work, we pursue this objective in two ways: (1) by introducing a new set of speech features, which we call them spectral energy distribution, or briefly SED. And (2) by reducing the dimensionality of the feature vector using the lasso. To see how optimal the proposed set of features are, we then compare the prediction performance of our work with those of the participants in the AVEC 2011.

The rest of this work is organized as follows. Section 2 is dedicated to the introduction of the new set of features. In Section 3, we review the linear family of models and the lasso. And the result of our experiment is presented in Section 4. This work will be brought to an end by concluding remarks.

2. Spectral Energy Distribution

The use of energy features for the recognition of emotional speech has been addressed before, in terms of the progression of energy in time [11], estimation of energy for different frequency bands [11, 12, 13, 3], total energy as a local feature [2, 14], and low and high frequency energy calculation [1]. In this work, we propose a set of features based on the spectral distribution of energy. Inspired by a work by Scherer and others [15], we call this set of features spectral energy distribution or briefly SED. We have recently proposed a similar set of features, to those as we are proposing in this work, and we called them FED (frequency-energy-distribution) [9]. Besides the name alteration, which is made in the favor of a more clear and conveying name and complying with the literature, there is a major difference between SED as we define in this work and FED as we have defined before. The difference is in the very heart of the definition of SED and that is the introduction of the function $g(.)$, as a normalizing function.

SED is composed of a set of components. Given the speech signal $s[n]$, we define component i as follows.

$$\text{SED}_s^i = \frac{2}{N}\sum_{k=1}^{N}(H[k-L_i]-H[k-U_i])|g(S[k])|^2. \quad (1)$$

Where $S[k]$ is the discrete Fourier transform (DFT) of $s[n]$ and $H[k]$ denotes the step function. L_i and U_i specify the lower and upper bounds of the component in the frequency domain. And N is the length of $s[n]$.

In this formulation, $g(.)$ is a normalizing function, defined as follows.

$$g(S[k]) = S^p[k],\ 0 < p \leq 1. \quad (2)$$

For the moment let's assume that $p = 1$. SED components are then no more than bins of the power spectrum of speech signal.

It is clear that different components have different shares in the total energy of the signal. That is to say, there are some components which take greater values than some others. However, this does not mean that components with greater values are of more importance in capturing the emotional contents of speech, than those with smaller values. Therefore, our intuition for incorporating the function g in the definition of SED is to normalize the components.

Depending on the size of each component (i.e. $U_i - L_i$), taking g out of the summation (Eq. 1) can also do the job, while saving on the computation.

3. Classification and Dimensionality Reduction

Given $(\mathbf{x}, y)$ pairs of input and output variables, analogously speech features and emotion labels, we are required to build a model that can convey the shared relationship among the pairs. Let's call such a model f and define it as follows.

$$\hat{y} = f(\mathbf{x}; \boldsymbol{\beta}). \quad (3)$$

Where $\boldsymbol{\beta}$ is a set of parameters, which specifies a model in a family of models described by f. $\hat{y}$ is the prediction of the model for a given instance like x. The hat on the variable name highlights the fact that for a given sample, the predicted value might be different than the actual value of the variable.

As for the choice of the family of models, we have chosen the linear family:

$$f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \beta_0. \quad (4)$$

Where $\mathbf{x}$ and $\boldsymbol{\beta}$ are vectors of size p and β_0 is a scaler. p is the dimensionality of the feature space.

The reason why we have made this choice is due to two reasons. One is that this family of models suggest a relatively simple description, which can be a plus for generalization considerations. The second reason is the strong literature [16] behind the linear models and the long developed techniques based on the family.

Let us now take a look at the choice of input variables. We know that not all the input variables $\{x_1, \ldots, x_p\}$ carry useful information. Dimensionality reduction is hence aimed at reducing the number of input variables, by ruling out redundant and misleading variables. There are two major approaches to dimensionality reduction: feature selection and extraction. While the result of feature selection is an explicit subset of primary features, feature extraction results in a combination of all the primary features. Looking at the literature [17], one may find forward selection, Fisher score, and information theoretic scores as some of many selection techniques. On the other hand, some examples of feature extraction methodologies are principal component analysis, Fisher discriminant analysis, and independent component analysis. In a search for an explicit set of features, we use a selection type of dimensionality reduction in this work.

As for the choice of feature selection, we choose the lasso. What lasso [18, 19] suggests is the regularization of learning criterion with ℓ_1 norm.

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \\ \text{subject to} \quad & \sum_{j=1}^{p}|\beta_j| \leq t. \end{aligned} \quad (5)$$

Where t is an adjusting parameter and N is the number of observations.

According to the lasso, those input variables that do not contribute to lessening of the squared error are penalized by a constraint. Therefore, the resulting $\boldsymbol{\beta}$ is sparse.

Those entries of $\boldsymbol{\beta}$ which are zero indicate the ruled out set of variables. In other words, the lasso combines the training of the parameters with a subset selection. The choice of t hence becomes substantial. An inappropriately large t may lead to very few non zeros entries of $\boldsymbol{\beta}$, or on the contrary, a small value of t may result in the selection of all the inputs. A proper choice of the parameter should be made through cross validation.

In the following section, we present some results of applying the proposed set of features, reduced by lasso, to emotional speech recognition.

4. Experimental Results

To verify the applicability of the proposed set of features to emotional speech recognition, we have done an experiment in the framework of the first international audio/visual emotion challenge (AVEC 2011) [3]. In this challenge, the solid-SAL part of the SEMAINE database [20] is used. As for the labels, four binary emotional dimensions are considered: activity, expectancy, power, and valence. For the sake of the consistency of the challenge, the database is provided in three parts: training, development, and testing. So, the contestants were asked to report their result on the testing set. Although we do not have access to the labels of the test set, we would like to make a comparison based on the prediction performance on the development set. The good thing here is that most of the participants have included the result of their algorithms on the development set in the papers, and that enables the possibility of a comparison. Therefore, in this experiment, we train our model using the training set and then apply it to the development set.

As of the extraction of SED, having performed a linear search, we have set the length of the spectral intervals equally to 100 Hz. Spectral intervals do not intersect and they cover 0 to 8 kHz (i.e. 80 intervals). The power of $g(.)$, as in Equation 2, is set to 0.2. Extraction is done in a local fashion, meaning that each of the features has been extracted from 100 mSec-long windows of signal. Statistics of the extracted features over the windows is then calculated. As for the choice of statistics, we have chosen minimum, maximum, mean, median, and standard deviation. The resulting feature vector is of a length 400.

The result of this experiment is shown on Table 1. For comparison purposes among participants in AVEC 2011, we have picked those papers which report their weighted average (WA) on the development set, rather than the unweighted average (UA). According to Table 1, the proposed model for emotional speech shows the best overall result in comparison to all the chosen participants, from both accuracy and complexity perspectives. From accuracy point of view, we can see the better performance of the proposed model for expectancy, power, and valence dimensions, as well as the mean WA over the four dimensions.

From complexity point of view, we can see how minimal the proposed model is. On the one hand, the proposed set of features includes not more than 15 features, whereas the length of feature vectors of other works ranges from 210 to more than 2500. On the other hand, this set of 15 features is composed of an explicit set of features described by SED. The proposed set of features come associates with 8 different spectral intervals of size 100 Hz. Table 2 gives a listing of the features and the intervals. According to this table, the use of the first spectral interval (SED^1) has seen for modeling expectancy, power, and valence dimensions. One may also notice that the intervals around 3 kHz have shown good capability in preserving emotional content of speech.

It is worthy to notice that Cen and others [5] have also used the lasso as for the choice of feature selection. However, they have run their study based on the baseline features of the challenge [3]. The baseline feature vector is a vector of a size 1941, composed of a variety features, including energy and voicing related, as well as spectral features. This can be due to two reasons. One is the sub-optimality of selection algorithms and that it becomes more severe as the size of a feature vector grows. Also, the informativeness of individual features in two feature vectors can contribute to this happening.

5. Concluding Remarks

Proposed in this work is a set of new speech features. We called them spectral energy distribution (SED) and defined them as a normalized binned power spectrum of speech signal. Our choice of modeling and dimensionality reduction are regularized linear models, specifically the lasso. To validate the applicability of SED to the emotional speech recognition, we applied it to the problem as considered in the first international audio/visual emotion challenge. When compared the prediction accuracy of the proposed model with those of the participants in the challenge, we noticed a better performance of the proposed model from two points of view. Firstly, our model results in a more accurate prediction, in terms of weighted average recall accuracy. Secondly, it takes as many as 15 explicit SED features to do the job, whereas the length of the feature vector of other participants ranges from 210 to more than 2500. We noticed that in spite of dimensionality reduction, a longer vector of features does not necessarily result in a better model.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, jan 2001.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

	Activation		Expectancy		Power		Valence		
	WA	NF	WA	NF	WA	NF	WA	NF	mean WA
Kim et al [7]	65.1	N/A	54.3	N/A	61.3	N/A	64.0	N/A	61.1
Calix et al [4]	63.8	1273	63.5	363	65.0	652	57.0	714	62.3
Schuller et al [3]	63.7	1941	63.2	1941	65.6	1941	58.1	1941	62.7
Cen et al [5]	58.7	≈1000	66.5	≈1000	65.9	≈1000	62.9	≈1000	63.5
Pan et al [8]	**65.4**	1941	66.5	621	64.5	1941	63.5	1941	65.0
Sayedelahl et al [9]	61.1	125	66.6	85	67.4	125	65.4	125	65.1
this work	61.3	**1**	**66.7**	**5**	**67.4**	**5**	**66.1**	**6**	**65.4**

Table 1: A comparison among some of the most recent works in the field, all performed in the framework of AVEC 2011. WA stands for weighted average (average of recall accuracy in percentage) and NF for no. of features used for the task.

Emotional Dimension	Activation	Expectancy	Power	Valence
Employed	mean SED[29]	max/mean SED[1]	mean/med SED[1]	mean SED[1]
Features		max/mean SED[14]	max/mean SED[28]	med SED[30]
		max SED[40]	mean SED[30]	max/med SED[32]
				mean/med SED[47]
Corresponding	2.8–2.9 kHz	0–0.1 kHz	0–0.1 kHz	0–0.1 kHz
Spectral		1.3–1.4 kHz	2.7–2.8 kHz	2.9–3 kHz
Intervals		3.9–4 kHz	2.9–3 kHz	3.1–3.2 kHz
				4.6–4.7 kHz

Table 2: The proposed set of features for modeling emotional contents of speech for each of the emotional dimensions, along with the corresponding spectral intervals.

[3] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 415–424.

[4] R. Calix, M. Khazaeli, L. Javadpour, and G. Knapp, "Dimensionality reduction and classification analysis on the audio section of the semaine database," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 323–331.

[5] L. Cen, Z. Yu, and M. Dong, "Speech emotion recognition system based on l1 regularized linear regression and decision fusion," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 332–340.

[6] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 359–368.

[7] J. Kim, H. Rao, and M. Clements, "Investigating the use of formant based features for detection of affective dimensions in speech," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 369–377.

[8] S. Pan, J. Tao, and Y. Li, "The casia audio emotion recognition method for audio/visual emotion challenge 2011," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 388–395.

[9] A. Sayedelahl, P. Fewzee, M. Kamel, and F. Karray, "Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 407–414.

[10] R. Sun and E. Moore, "Investigating glottal parameters and teager energy operators in emotion recognition," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 425–434.

[11] K. Ślot, J. Cichosz, and L. Bronakowski, "Emotion recognition with poincare mapping of voiced-speech segments of utterances," in *Artificial Intelligence and Soft Computing – ICAISC 2008*, ser. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, vol. 5097, pp. 886–895.

[12] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157–183, 2003.

[13] L. ten Bosch, "Emotions, speech and the asr framework," *Speech Communication*, vol. 40, no. 1-2, pp. 213–225, 2003.

[14] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition," *Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[15] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," in *Handbook of Affective Sciences*, ser. Series in Affective Science, Oxford University Press, 2003, pp. 433–456.

[16] G. Seber and A. Lee, *Linear Regression Analysis*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2012.

[17] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[18] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer series in statistics. Springer, 2008.

[19] R. Tibshirani, T. Hastie, and J. H. Friedman, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 01, 2010.

[20] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, july 2010, pp. 1079 – 1084.