Features Importance Analysis for Emotional Speech Classification

Jianhua Tao and Yongguo Kang

National Laboratory of Pattern Recognition (NLPR), Institute of Automation,
Chinese Academy of Sciences, P.O.X. 2728, Beijing 100080
{jhtao, ygkang}@nlpr.ia.ac.cn

Abstract. The paper analyzes the prosody features, which includes the intonation, speaking rate, intensity, based on classified emotional speech. As an important feature of voice quality, voice source are also deduced for analysis. With the analysis results above, the paper creates both a CART model and a weight decay neural network model to find acoustic importance towards the emotional speech classification and to disclose whether there is an underlying consistency between acoustic features and speech emotion. The result shows the proposed method can obtain the importance of each acoustic feature through its weight for emotional speech classification and further improve the emotional speech classification.

1 Introduction

More and more efforts have been made for the research of emotional speech recently. It is a widely known fact that the emotional speech differs with respect to the acoustic features [2]. Some prosody features, such as pitch variables (F0 level, range, contour, and jitter), speaking rate are analyzed [2]. Parameters describing laryngeal processes on voice quality were also taken into account [3]. The method of using acoustic prosodic cues to classify emotions or speaking style has been adopted by many researchers. Huber used 50 neutral and 50 angry utterances and multi-layer perceptions for classification. Results reached around 90% of accuracy in the simplified tasks of distinguishing emotional from non-emotional utterances. Dellaert *et al.* [8] compared three classifiers: the maximum likelihood Bayes classification, kernel regression, and k-nearest neighbor (K-NN) method with particular interest in sadness, anger, happiness, and fear. They used features from the pitch contour. An accuracy of 60%~65% was achieved. Valery A. Petrushin[9] performed an experimental study on vocal emotions and the development of a computer agent for emotion recognition. Noam Amir[10] used a corpus that had been studied extensively through subjective listening tests. Best results obtained using distance measures based classifiers. Lee *et al.*[11] used linear discriminant classification with Gaussian class-conditional probability distribution and K-NN method to classify utterances into two basic emotion states, negative and non-negative. Tato *et al.* [12] discussed techniques that exploit emotional dimension other than prosody. Their experiments showed how “quality features” were used in addition to “prosody features.” Yu *et al.* [13] used SVMs for emotion detection. An average accuracy of 73% was reported.

J. Tao, T. Tan, and R.W. Picard (Eds.): ACII 2005, LNCS 3784, pp. 449 – 457, 2005.

In the paper, to know how the acoustic features affect the final perception results, we did some more analysis about feature importance for emotional speech classification. We have built both a CART model and a weight decay neural network to learn the relationships between the acoustics and the perceptual characteristics. Based on this, the importance of acoustic features was analyzed.

The whole paper is broken down into several major parts. Section 2 introduces the corpus which supports all of the later work in the paper. Furthermore, the paper makes some analysis on acoustic features (F0 mean, F0 top, F0 bottom, speaking rate, voice source, etc.) according to the different emotion states. In section 3, the paper creates a CART to do the emotion classification based on above acoustic parameters. The acoustic importance is analyzed according to the training results. In section 4, the paper describes a weight decay neural network which is used for automatic analysis of feature importance. Finally, the feature importance analysis results from two models are compared and discussed in section 5. Some other factors which might influence the perception will also be discussed here.

2 Corpora and Statistic Results of Acoustic Features

The corpora used for analysis contains 1,2000 sentences (8 hours), which are performed by 2 actors and 2 actresses with 5 emotions, neutral, fear, sad, angry, happy. They are segmentally and prosodically annotated with break index, F0 contours, syllable boundaries, transcription, etc. All of the results are manually checked. From previous research [2], we know F0, speaking rates, intensity and voice quality form the important features for emotion classification. To know the behavior of each parameter in different emotion states, we made more experiment on the parameters' distributions. The results are shown in Fig.1 and Fig.2. In both Fig.1 and Fig.2, the X-coordinate means the different type of emotion states, from neutral to happy. Y-coordinate denotes the mean value (dots) and standard deviation (vertical lines) of each parameter within the single emotion state.

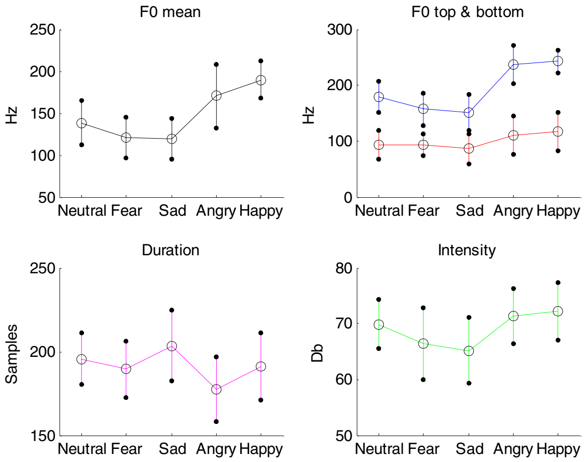

Fig. 1. Relationship between normalized F0s (top), syllable duration (bottom left), intensity (bottom right) and emotion states

Fig.1 shows the distribution of prosody features. In the figure, it shows that “happy” and “angry” make a very high F0, while “sad” generates lower value than neutral state. F0 parameters of “fear” make quite similar behaviors as “sad”. “sad” utterances are normally slower than those of other emotions, while “angry” is the fast. “Angry” and “happy” have relatively higher power than other emotions, while “fear” and “sad” are the lowest. All of the results confirm the previous research results [6].

From Fig.1, we also can find the overlap of F0 mean and F0 top in different emotions is less than that of F0 bottom. That means F0 mean and F0 top seem to be the better “resolving power” for perception than F0 bottom. It gives us a brief impression on which is might be more important in distinguishing the emotional speech.

Normally, we think voice quality is a kind of timbre for speech, and is determined by laryngeal characteristics. There are some kinds of phonation modes, such as breathy, whisper, falsetto, creaky, normal, and so on, which correspond to certain laryngeal characteristics respectively. However, there would be some subtypes within one category. In normal mode, one end of the continuum of subtypes approaching breathy voice, where the laryngeal muscles controlling vocal fold adduction are relatively relaxed. At the other end, tension in the musculature begins to limit the vibration of the folds and voice verges on laryngealized or creaky voice [6, 7].

Usual acoustic measurements of breathiness could be classified into two main categories of source parameters and spectral parameters according to the estimate procedure [6]. A general source model is a four-parameter liljencrants-fant(LF) model [5], whose parameters are Ee (the excitation strength), Ra (the measure of the return phase), Rk (the measure of the symmetry/asymmetry of the glottal pulse), and Rg (the measure of the opening branch of the glottal pulse). The familiar parameter, open quotient (Oq), is defined as (1+Rk)/2Rg. The statistic results of the voice source parameters are shown in Fig.2.

From Fig.2, we can see that the mean values of “sad” utterances’ glottal sources are distinctly less than those of other emotional utterances'. The low Ee of “sad” utterances indicates its overall weak source signal as well as whispery voice. Consistent with Johnstone's experiments [7], extremely low Ra values indicate a sharp instantaneous closure of the glottis. In this regard, “sad” utterances, which are likely to reflect the relatively high degree of laryngeal tension, are similar to tense voice [10]. Thus “sad” utterances have the trend of exhibiting whispery and tense voice qualities. However, compared to higher Sq of tense voice, the mean value of Sq of “sad” utterances is lower too. The wildly suggested association of “angry” emotion with tense voice is not supported in this study. Although the mean values are employed in the current research, it can be indicated that the relation between emotion and voice quality is not one to one and an emotional speech can exhibit several characteristics of different voice qualities. In other words, concluded by Gobl [3], a given quality tends to be associated with a cluster of affective attributes.

Corresponding with their range, it can also be observed from the figures that standard deviations of Sq are the largest and standard deviations of Ra are the smallest among the six glottal parameters. Comparing the five emotional states, “angry” utterances have the largest standard deviation for Ee and Sq, while “sad” utterances have the largest standard deviation for other four parameters. The standard deviation represents the distribution of parameters, so it can be indicated that the variation of spectral intensity is the largest in “angry” utterances while the variation of voice quality is largest in “sad” utterances.

As mentioned above, Ee is the absolute value of the negative peak of the differentiated glottal flow and is correlated with the overall intensity of the signal. Generally speaking, “angry” utterances sound louder than those of other emotions and should have larger Ee too. So, larger Ee is not critical for producing “angry” utterances, in particular “angry” female utterances.

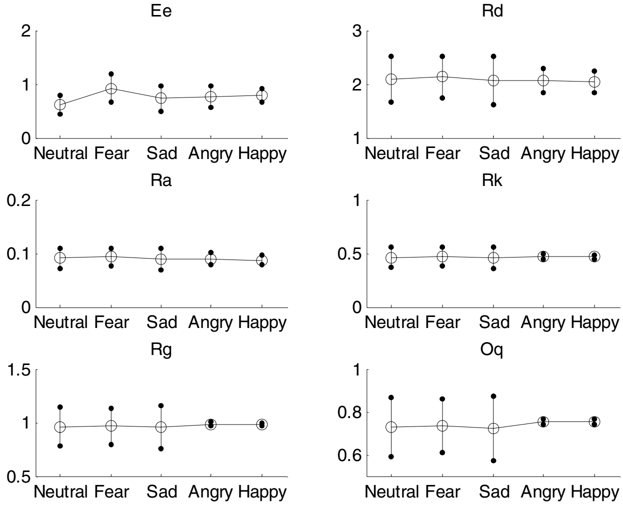

Fig. 2. Voice source properties among different emotions

Previous work has confirmed the following features to be useful for emotional speech classification: utterance duration, f0-range, f0-variation, f0-maximum, f0-minimum, f0-mean, power-level, power-mean and voice source parameters. To simulate the stress information, we add some more parameters, position of f0 peak in the utterance, position of f0-minimum in the utterance, position of duration peak in the utterance, position of the duration minimum in the utterance. After we got the general idea of the influence between the emotional speech and the groups of acoustic features, we still wish to know more subtle acoustic influence of each acoustic parameter on emotions. In next two sections, classification and regression tree (CART) and weight decay neural network methods are employed to analyze the importance of every acoustic feature.

3 Acoustic Importance Analysis with CART Model

This section, we built a CART model to learn the relationships between the acoustics and the emotion states in order to predict the most likely response for each speech token for a reclassification. We used simple first-order statistics derived from the acoustics as the independent variables. The tree correctly predicted 69% of categories using 26 leaf nodes. After training, we used a function that passed the same acoustic data to each of the classification trees, and allowed each to output a probability indicating the likelihood of a sub-parameter of expressiveness vector. These likelihoods were ranked and an ordered list of relevant categories was produced for each speech token.

To know how the acoustic features affect the final stimuli perception, we made further analysis. As we know, variable importance, for a particular predictor, is the sum across all nodes in the tree of the improvement scores that the predictor has when it acts as a primary or surrogate (but not competitor) splitter. Specifically, for node i, if the predictor appears as the primary splitter, then it has a contribution toward the importance as:

importance_contribution_node_i = improvement

If, instead, the predictor appears as the n'th surrogate instead of the primary predictor, the expression is:

*importance_contribution_node_i = (p ^ n) * improvement*

in which p is the "surrogate improvement weight": a user controlled parameter which is equal to 1.0 by default and can be set anywhere between 0 and 1. Linear combination splits do not contribute in any way to variable improvement.

If, in the absence of linear combinations, the improvement weight is greater than 0, and the variable has importance = 0.0, it does not appear in the tree as a primary or surrogate splitter, although it may appear as a competitor.

With this method, we got the factors related to each input acoustic parameters after the training in Table 1.

Table 1. Ranking score of input acoustic parameters

Parameters	Ranking Score
F0 mean	100.00
F0-maximum	98.79
F0-range	98.64
Ee	57.63
Duration mean	48.34
Duration Range	38.78
Position of F0 minimum	36.36
Power	34.86
Rd	33.52
F0 minimum	33.46
Ra	30.31
Position of F0 maximum	29.86
Rk	27.12
Rg	23.40
Power range	15.07
Position of maximum duration	14.02
Oq	13.18
Position of minimum duration	1.56

The parameters, whose ranking score are zero, are not listed in the table. It is easy to find that the F0 mean assumes the most important role in emotion perception. Ee is the most important parameter related to voice quality for the model. Position of F0 minimum is, then, the most important word stress feature for emotion perception.

4 Acoustic Importance Analysis with Weight Decay Neural Network

The weight decay concept is well known in neural network theory as a type of regularization [1]. Regularization is typically used to reduce the complexity of a NN by adding a penalty term $P(w)$ to the error function $F(w)$:

$$\overline{F}(w) = F(w) + \lambda \cdot P(w) \tag{1}$$

where w denotes a vector containing all weights in the NN and λ controls the influence of the penalty term. In the scope of this work $P(w) = \sum_i w_i^2$, which is known as standard weight decay (i=1, …, number of weights). λ is then typically referred to as decay rate.

During training, weights will be adopted on the basis of this function (using gradient descent):

$$w^{i+1} = w^i - \eta \nabla \overline{F}(w) \tag{2}$$

The parameter η is generally referred to as learning rate and controls the step size used to adapt the weights. In our case η and λ are kept constant in all steps i.

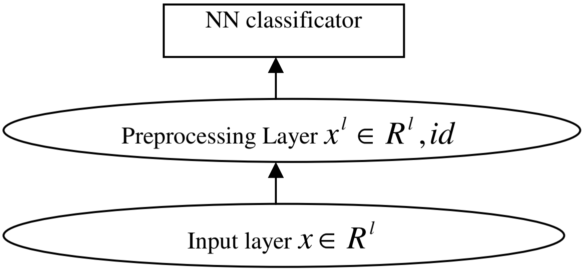

Fig. 3. NN with preprocessing layer

To find out which of the input parameters are important to the NN for a specific task, a preprocessing layer is inserted between the input and the network. The input signals $x \in R^l$ are propagated to this preprocessing layer via a diagonal matrix $w_{diag} = diag(w_1, ..., w_l)$ to give the output signals $x^l \in R^l$ of the preprocessing layer. Figure 3 shows the resulting network architecture. The weight decay concept is applied only to weights of the diagonal matrix. Thus, w_{set} contains all elements of

$w_{set} = w_l = \{w_1, \ldots, w_l)$. For the neurons of the preprocessing layer, the identity function is chosen as activation function.

At the beginning of the training process all elements of the diagonal matrix are initialized with 1. Thus, the input signals are transferred to the hidden layer without modification. In each training epoch i eq. (3) is applied to obtain the new values for epoch (i+1) for the weights in the diagonal matrix. It is important to carefully choose the decay rate λ. λ should generally be as small as possible. This way the influence of the learning rate η on weight adaptation (eq.(3)) is stronger than the influence of the decay rate λ. Therefore, non-linear relations hidden in the data can be captured. On the other hand λ should be large enough, so that it effects the weights in the diagonal matrix. After several training epochs and application of eq.(3) to the weights in w_l, the following behavior is observed: For some elements of w_l the influence of the learning rate η is stronger than the influence of the decay rate λ. For other elements of w_l, however, the influence of the decay rate λ is stronger than the influence of the learning rate η. By choosing λ / η right, some weights can be pushed towards zero, while others range higher. Those weights close to zero, or below a certain threshold, are considered to be of less importance to the training success of the auto-associator classificatory network. All weights of the auto-associator classificatory network are trained without the penalty term $P(w)$ of (1) at the same time as the weights in w_l.

The table 2 shows the value w_l, Some relatively high values can clearly be identified.

Again, the parameters whose ranking score is zero are not listed in the table. Similar to the results deduced from CART model, F0 mean also plays the most important role in emotion perception. Ee is the most important parameter related to voice quality. Unlike the previous results, position of F0 maximum seems to be the most important word stress feature with NN model.

Table 2. w_l of input parameters

Parameters	w_l
mean F0	0.51
maximum F0	0.42
Ee	0.38
mean Duration	0.36
position of maximum F0	0.35
Intensity	0.35
Rd	0.32
minimum F0	0.28
Ra	0.25
Rk	0.25
Rg	0.24
Oq	0.21

5 Discussion

In the sections above, two methods including CART and neural network are used to explore the importance of each acoustic feature. Although the results of two methods are not entirely identical, the underlying relationship between acoustic features and emotional states is disclosed to a certain extent. Observed from these results, The F0 features such as f0 mean are most important features, while duration, intensity and voice source features are useful features for emotional speech classification in turn.

The above analyses are just based on the one person's speech. As we know, there are different properties of acoustic features among the speakers, even though there is no difference in emotion expression. To show how the speech differs from the different persons we make some more statistic results of parameter Rg among five people. It is shown in Fig.3.

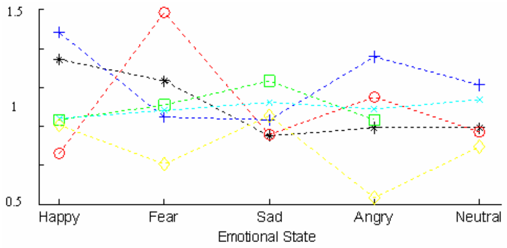

Fig. 3. Rg distribution among five different people

We see Rg is quite unstable among the persons. It may be helpful for speaker identification system, but may be not suitable for creating a speaker independent emotion classification system. Similar analysis of other parameters will be done in our future work. The importance order of the acoustic parameters might be changed, while using the speech from several persons.

6 Conclusion

In this paper we make some analysis on the acoustic parameters, especially the voice source parameters, in different emotional speech. Based on that, classification tree and the weight decay concept known from neural network theory have been applied in the emotion speech classification. By applying above two methods, the importance of acoustic parameters could be evaluated. The results will be much helpful for the later research on the emotion classification and emotional speech synthesis.

References

[1] Nick Campbell, "Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation", ICSLP2004, Jeju, Oct, 2004.

[2] Jianhua Tao, "Emotion control of chinese speech synthesis in natural environment," in EUROSPEECH- 2003, pp. 2349–2352.

[3] C. Gobl and A. N´ı Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” Speech Communication, vol. 40, pp. 189–212, 2003.

[4] Scherer K.R., “Vocal affect expression: A review and a model for future research,” Psychological Bulletin, vol. 99, pp. 143–165, 1986.

[5] G. Fant, Liljencrants J., and Q. Lin, “A four- arameter model of glottal flow,” STL-QPSR 4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 1–13, 1985.

[6] Yildirim, Serdar Bulut, Murtaza Lee, Chul Min Kazemzadeh, Abe Deng, Zhigang Lee, Sungbok Narayanan, Shrikanth Busso, Carlos (2004): "An acoustic study of emotions expressed in speech", In INTERSPEECH-2004, 2193-2196.

[7] T. Johnstone and K. R. Scherer, “The effects of emotions on voice quality,” in Proceedings of the XIVth International Congress of Phonetic Sciences, 1999.

[8] Dellaert, F., Polzin, t., and Waibel, A., Recognizing Emotion in Speech”, In Proc. Of ICSLP 1996, Philadelphia, PA, pp. 1970-1973, 1996.

[9] Petrushin, V. A., “Emotion Recognition in Speech Signal: Experimental Study, Development and Application.” ICSLP 2000, Beijing.

[10] Amir, N., “Classifying emotions in speech: a comparison of methods”. Holon Academic Institute of technology, EUROSPEECH 2001, Escandinavia.

[11] Lee, C.M.; Narayanan, S.; Pieraccini, R., Recognition of Negative Emotion in the Human Speech Signals, Workshop on Auto. Speech Recognition and Understanding, Dec 2001.

[12] Tato, R., Santos, R., Kompe, R., Pardo, J.M., Emotional Space Improves Emotion Recognition, in Proc. Of ICSLP-2002, Denver, Colorado, September 2002.

[13] Yu, F., Chang, E., Xu, Y.Q., and Shum, H.Y., Emotion Detection From Speech To Enrich Multimedia Content, in the second IEEE Pacific-Rim Conference on Multimedia, October 24-26, 2001, Beijing, China.