The Friendliness Perception of Dialogue Speech*

Jianhua Tao Lixing Huang Yongguo Kang Jian Yu

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing
{jhtao, lxhuang, ygkang, jyu}@nlpr.ia.ac.cn

Abstract

The paper is focused on the friendliness analysis and perception of dialogue speech. To do that, the paper uses a concept of the "perception vector" which contains the information of emotions and softness. In creating the "perception vector", and to simulate the perception ambiguity, the paper allows the listeners to label the speech with multiple emotions, and align them into "one choice", "first choice" and "second choice". Then, the paper makes the correlation analysis between friendliness and "perception vectors", the results disclose that the friendliness is positive correlation to "softness", "happiness" and "anger". Finally the paper traines a classification tree model to predict friendliness degree from acoustic features. With the classification tree model, we get the ranking scores of the acoustic parameters' importance for perceptually synthesized speech. Results shows that the F0 mean assumes the most important role in emotion perception, Ee is the most important parameter related to voice quality for the perception model.

1. Introduction

Recently, more and more efforts have been made for the research of affect speech. Aomg them, conversational speech is one of most popular communication methods between human and human. It contains more spontaneous and para-linguistic information. In application, especially in service center, many people are interested if customers are satisfied with their service or not. Do they make complaint, anger or joy with the service? In the paper, we, then, use "friendliness" to represent this spectrum of valenced feeling states and attitudes for it, with positive friendliness representing the pleasant or polite end (e.g., feeling grateful,; expressing with soft moods) and "negative friendliness" representing the unpleasant and un-polite end (e.g., feeling contemptuous, irritable; expressing with hard moods). The affective texture of a person's life— or of a given relationship or group—can be represented by its "friendliness degree", the degree of friendly feelings.

In the paper, we selected five basic emotions ("anger", "happiness", "sadness", "fear" and "neutralness") as the reference states for analysis. Since most of the dialogue speech is closed to "neutral" states. We introduced "softness" to be another feature to represent if the customers/operators are polite or not. Softness is a sound property that is free from loudness or stridency. All of these features were used to generate a "perception vector". In order to get the fuzzy perception of emotions, each utterance was allowed to be labelled with more than one emotion state. They were aligned with the "one choice", "first choice" or "second choice". In the paper we did more analysis in which we found that the "softness", "happiness" and "anger" are positive correlation to friendliness degree. The high "softness degree" and high "happiness" with low "anger" usually describe a friendly utterance. On the contrary, the high "anger" with lower "softness degree" and low "happiness" denotes a very low "friendliness degree".

Furthermore, the paper builds a CART model to learn the relationships between the acoustics parameters and the friendliness degree of the utterance. Based on the model, the paper make an importance analysis for the acoustic paramerters. The analysis results disclose that F0 features are still the most important parameters in determining the friendliness of the speech.

The whole paper is broken down into five major parts. Section 2 introduces the corpus preparing, in which each sentence of one speaker was labelled with friendliness degrees, softness degrees and emotions. In emotion labeling, the multiple selection of emotion states is allowed. Section 3 generates a perception vector from emotions and softness. The vector is, then, used for the analysis of the relationship among friendliness, emotions and softness. In Section 4, the paper builds a CART model to learn the relationship between acoustic features and friendliness degree. The acoustic importance analysis is also done here. Section 5 makes some discussion and final conclusion of the paper.

2. Corpus and labelling

2.1. Corpus preparing

The corpus used in the paper is collected from a call center via ten telephone lines. The SNR is various from 20db to 30db. The sampling rate is 16kHZ. The resulting corpora contains about 20 hours with 26 peoples and 10 operators. To avoid the mixture of the speech during the conversation, each channel was recorded separatedly, and each speech file contains only one speaker. The topics are limited in service consulting. After recording, the corpus was separated into sentence by sentence with the transcription. All of them are segmentally with syllable boundaries and prosodically annotated with F0 values.

2.2. Labeling

When corpus was segmented into sentence level, we performed a perception experiment, in that we excised utterances with a random list to a group of 10 subjects. The listeners were graduated students of the university with a grade-point for their cooperation in the experiment. The perception was performed in three steps.

Step 1: Label the friendliness degree from 0 to 10 for each utterance (each utterance contains only one speaker). If the listener felt the speaker was willing to speak, explain, chat, …, with a very polite or pleasant way, he/she could give a high friendliness degree. On the contrary, the un-polite or angry way, etc. will result in a lower "friendliness degree".

* The paper is supported by National Natural Science Foundation of China (No. 60575032)

Step 2 (two months later after the first step): Label the softness degree from 0 to 10 for each utterance by the same listeners of the first step. The listeners were asked to label it according to their feeling in pitch and voice quality. The pleasant feeling with soft pitch and voice timbre will give higher softness degree.

Step 3 (two months later after the second step): Label the possible emotion states of each utterance by the same listeners of the above two steps. In labeling the emotions, we met a problem that it is hard for a listener to determine emotion states from the dialogue speech which doesn't contain strong moods in the most parts. Even for strong moods, the speech might still cause different perception results. Someone thinks it as "happiness," but others might consider it as "neutralness." To avoid this problem, the simplified method is asking the listeners labeling the degrees for each emotion they perceived. But there was too much speech to be labelled, instead, the listeners were asked to note the emotion with one or two states from a list of "happiness, fear, sadness, anger and neutralness." To show which is the better choice, the alignment of the selections are also required.

The listeners have considerable freedom in their choice of labels. If the listener has strong feeling that the sentence is related to an emotion state, just one state is selected, otherwise, he/she is asked to select two states and line them with "first choice" and "second choice" according to the comparison between two selection results. Different listeners perceive different aspects of this multi-faceted phenomenon and it can be difficult to achieve a consensus on the choice of a single most appropriate label for any given speech utterance.

Here shows some response counts from one listener:

utt 1: happiness, neutralness, softness(5), friendliness(7)
utt 2: neutralness, softness(8) , friendliness(7)
utt 3: neutralness, softness(7) , friendliness(8)
utt 4: sadness, fear, softness(7) , friendliness(8)
utt 5: neutralness, softness(7) , friendliness(7)
utt 21: sadness, neutralness, softness(6) , friendliness(7)
utt 22: anger, softness(2) , friendliness(2)
utt 23: anger, softness(2) , friendliness(2)
utt 24: neutralness, softness(5) , friendliness(5)
utt 25: neutralness, softness(7) , friendliness(7)

From the results, utterance 1, for example, is rated in "happiness" and "neutralness." It means the listener thinks this sentence might be both "happiness" and "neutralness." But "happiness" seems to be stronger than "neutralness." The softness degree of the utterance is labelled as "5" and the friendliness degree is 7. Utterance 22 has only one labeling result, "anger." It shows the listener can make sure the decision.

3. Correlation Analysis

3.1. Generating Perception Vectors

The experiment, above, has helped us to understand the complexity of perception, especially in a spoken utterance. As a result, we believe that it is hard to label the speech with one-right-answer, and that it is better to represent this type of paralinguistic information by using a vector of probabilities instead. Underlying the multiple perception selection, we build a vector of probabilities in a set of "happiness," "sadness," "fear," "anger", "neutralness" and "softness degree", for each possible response. We hope the friendliness can be deduced from it.

We select utterance 1 as an example. The perceptual results of the utterance from all listeners are shown in the following table.

Table 1, Perceptual results of utterance 1from all listeners

Listener ID	One/First Choice	Second Choice	Softness Degree	Friendliness Degree
1	happiness	neutralness	5	7
2	happiness		5	6
3	neutralness	happiness	5	7
4	neutralness		6	6
5	happiness		6	6
6	happiness		4	7
7	happiness	neutralness	4	7
8	neutralness		5	6
9	neutralness	happiness	3	4
10	neutralness		4	5

For "one choice" result, we assign the corresponding emotion state as a weight 1.0, the weight of the "first choice" is α, the second choice is $1-\alpha$. The perception of this utterance is formed by the weights vector of the five basic emotions and softness degrees. A vector (n, f, s, a, h, t) is used to denote it. The parameters in the vector denotes the probability of "neutralness", "fear", "sadness", "anger", "happiness". t is the normalized softness degree (from 0 to 1), which was got by

$$t = (\frac{1}{N}\sum_{n=1}^{N} t_n)/10 \quad (1)$$

Here, t_n denotes the softness degree labelled by the listener. N is the number of the listeners who took part in the perception experiments. Here, $N=10$. According to the table 1, $t = (\frac{1}{10}(5+5+5+6+6+4+4+5+3+4))/10 = 0.47$

The other parameters were got with the following methods,

$$n = \frac{1}{N}\sum_{i=1}^{N}\omega_{n,i}, \qquad f = \frac{1}{N}\sum_{i=1}^{N}\omega_{f,i}, \qquad s = \frac{1}{N}\sum_{i=1}^{N}\omega_{s,i}$$

$$a = \frac{1}{N}\sum_{i=1}^{N}\omega_{a,i}, \qquad h = \frac{1}{N}\sum_{i=1}^{N}\omega_{h,i} \quad (2)$$

$\omega_{n,i}$, $\omega_{f,i}$, $\omega_{s,i}$, $\omega_{a,i}$, and $\omega_{h,i}$ denote the weights of emotion states – "neutralness," "happiness," "sadness," "anger" and "fear", labelled by listener i.

For instance, if i=1, then $\omega_{s,i}=0$, because listener 1 didn't assign "sadness" for the utterance. The other parameters are, $\omega_{h,i}=\alpha$, $\omega_{a,i}=0$, $\omega_{f,i}=0$, $\omega_{n,i}=1-\alpha$.

For all 10 listeners, there are two "neutralness" and two "happiness" for the "second choice", two "happiness" and one "neutralness" for the "first choice", three "happiness" and three "neutralness" for the "one choice".

So, the final parameters for the vector are,

$$n = \frac{2(1-\alpha)+\alpha+3}{N} = \frac{5-\alpha}{10},$$

$$h = \frac{2(1-\alpha)+2\alpha+3}{N} = \frac{5}{10} = 0.5,$$

$s=0$, $a=0$ and $f=0$ (3)

The perception of the utterance is got,

$$(n,f,s,a,h,t) = (\tfrac{5-\alpha}{10},0,0,0.5,0) \quad (4)$$

To simplify the analysis, the "friendliness degree" is also normalized by the mean results from all listeners:

$$P' = (\frac{1}{N}\sum_{n=1}^{N} p_n)/10 \quad (5)$$

Where p_n denotes the softness degree labelled by the listener. The results were unified into the space [0, 1].

3.2. Correlation analysis between friendliness and perception vectors

For all of the later analysis, we use an assumption that $\alpha = 0.7$. We selected three typical status of normalized friendliness degree for analysis. Their distributions are limited into the following spaces, [0.7, 0,9], [0.4, 0.6] and [0.1, 0,3]. We, then, draw the statistic results of perception vectors in figure 1, 2 and 3. The X-coordinate means the parameters of the vector. Y-coordinate denotes probability of each parameter. Different lines show their standard deviation, while the dots mean the mean values of each parameter.

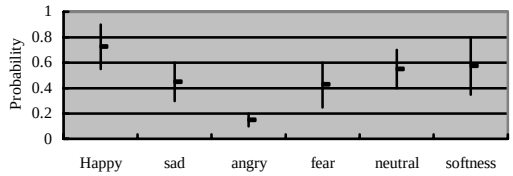

Figure 1, the statistic perception vector distribution of the normalized friendliness degree from “0.7” to “0.9”

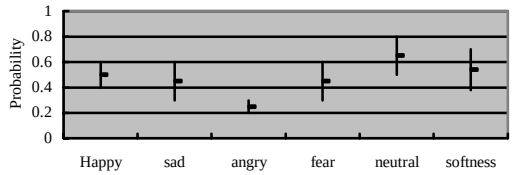

Figure 2, the statistic perception vector distribution of the normalized friendliness degree from “0.4” to “0.6”

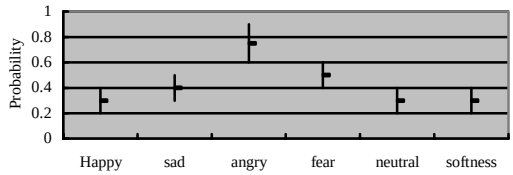

Figure 3, the statistic perception vector distribution of the normalized friendliness degree from “0.1” to “0.3”

From all of the parameters, we can find that friendliness is positive correlated to the “softness”, “happiness” and “anger”. The high “softness degree” and high “happiness” with low “anger” usually describe a friendly utterance. On the contrary, the high “anger” with lower “softness degree” and low “happiness” denotes a very low “friendliness degree”. The “sad” and “fear” contain a large range of distribution among most of friendliness degrees. It discloses the fact that there are big arguments in labeling the friendliness while the listener felt a “sadness” or “fear”. It is also hard to find the rules for these two states.

3.3. Correlation analysis between softness and emotions

From the above figures, we see the softness behaves a very important feature for friendliness perception, especially in weak emotions or neutral state. To know the relationship between the softness and other emotions, we made another experiment, in which we selected three distribution of normalized softness degree, [0.7, 0,9], [0.4, 0.6] and [0.1, 0,3], and draw the statistic results of emotions in figure 4, 5 and 6.

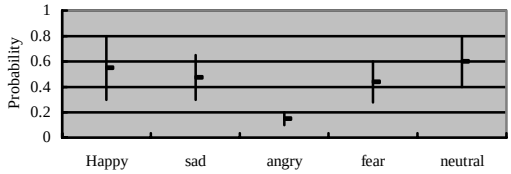

Figure 4, the statistic emotion distributions of the normalized softness degree from “0.7” to “0.9”

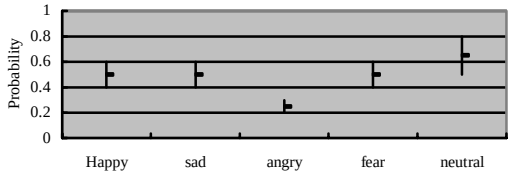

Figure 5, the statistic emotion distributions of the normalized softness degree from “0.4” to “0.6”

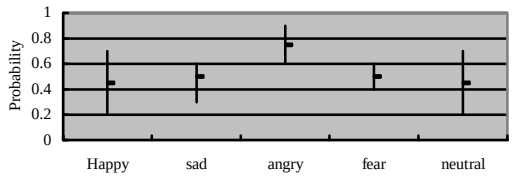

Figure 6, the statistic emotion distributions of the normalized softness degree from “0.1” to “0.3”

From figures, we find the strong “anger” is always related to lower softness. It means “anger” is the important factor for both friendliness and softness. It is interesting to find that the emotion distributions of softness are very similar to friendliness, while the normailized softness degree is higher than 0.5. So, we can say most of the higher friendliness degree can be determined by softness character. In lower softness, we find the emotion probabilities are various in a big range except “anger”. So, the “anger” might to be the better factor to determine the lower friendliness than the softness.

4. Correlation Analysis from Acoustic Features

4.1. Acoustic features

Many research results have proved that affect speech differs with respect to the acoustic features. Some prosody features, such as pitch variables (F0 level, range, contour, and jitter), speaking rate have been analyzed [6].

Parameters describing laryngeal processes on voice quality have also been taken into account [3]. 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 [3]. In our paper, we select voice source parameters to represent the voice quality features approximately, and select the utterance duration, f0-range,

f0-variation, f0-maximum, f0-minimum, f0-mean, power-level, power-mean for prosody features. A general source model is a four-parameter Liljencrants-Fant(LF) model, 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. It has been found that breathy voice has high Ra, Rk, and Oq values [4].

4.2. CART model

The method of using acoustic prosodic cues to classify emotions or speaking style has been adopted by many researchers with different methods, such as multi-layer perceptions based method[5], the maximum likelihood Bayes method[6], the k-nearest neighbor (K-NN) method[7], the distance measures based classifiers [9], the linear discriminant classification method, and SVMs method, etc..

In the paper, 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 28 leaf nodes.

4.3. Analysis

In CART model, 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. From the results, we found 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 maximum is, then, the most important stress feature for emotion perception.

5. Discussion and Conclusion

When analyzing the affect speech, we are easy to fall into the practice of processing the emotional speech. Actually, the most of speech around us contain much more information then just emotions. For dialogue speech, the "friendliness" is one of the most interesting features, it is related to emotions, but much more than that.

The paper generates a perception vector, which contains emotions and softness, to simulate the status of friendliness. To get the perception ambiguity, the paper described a perception experiment that allow us to label the speech with multiple emotions with degrees of "one choice", "first choice" and "second choice". The correlation analysis between friendliness and perception vectors discloses that the friendliness is positive correlation to "softness", "happiness" and "anger".

Finally the paper trained a classification tree model to predict the friendliness degree from acoustic features derived from the speech tokens. With the classification tree model, we get the importance of the acoustic parameters for friendliness perception. The results are helpful for the later research.

However, both friendliness and softness are complicated to be labelled, they are influenced by many psychological factors. Normally, it needs lots of subjects to label the corpus to ensure the statistical analysis. Unfortunately, we only could get 10 listeners. It took them two weeks to finish one labeling step. Even for that, we still think the analysis results of correlation between "friendliness" and "emotions and softness" are still reasonable.

The second factor which might influence the research is the noise of the voice. We've tried to select the speech with smaller noise, but still cannot ensure the recording quality, since all of the speech was recorded via telephone lines, and some speakers were out of the office. With that, some acoustic features cannot be acquired reliably.

Including the above reasons, the different characters of speakers and different speaking styles might also influence the perception results and make some errors in friendliness degree prediction from the acoustic features.

However we have lots of limitation of the experiment in the paper, we believe the research results will still be a good reference work for the later research.

6. 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, Yongguo Kang, "Features importance analysis for emotional speech classification," ACII2005, 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] 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.

[5] 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.

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

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

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

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