INTERSPEECH 2009 BRIGHTON

Detecting changes in speech expressiveness in participants of a radio program

Plínio A. Barbosa

Speech Prosody Studies Group/Dep. of Linguistics/Inst.Est. Ling., Univ. of Campinas, Brazil

pabarbosa.unicampbr@gmail.com

Abstract

A method for speech expressiveness change detection is presented which combines a dimensional analysis of speech expression, a Principal Component Analysis technique, as well as multiple regression analysis. From the three inferred rates of activation, valence, and involvement, two PCA-factors explain 97 % of the variance of the judges' evaluations of a corpus of radio show interaction. The multiple regression analysis predicted the values of the two listener-oriented, PCA-derived dimensions of promptness and empathy from the acoustic parameters automatically obtained from a set of 206 utterances produced by radio show's participants. Analysed chronologically, the utterances reveal expression change from automatic acoustic analysis.

Index Terms: expressive speech, appraisal inference, expression dynamics

10.21437/Interspeech.2009-615

1. Introduction

The pervasiveness of human expression in everyday, audiovisual interactions is undeniable. The need for collecting data in natural conditions of communication directly follows the interest of mainstream research on speech prosody in studying natural communicative functions, and, for this purpose, to find alternatives to lab speech [8, 5] and actor portrayed speech [6]. The use of corpora containing speech samples of natural affective states occurring during TV and radio shows [12], computer game plays [6], map and object-based tasks [7] is thus ideal for the analysis of expressive speech. According to Scherer [12], the main drawbacks of this kind of material are threefold: short utterances, noisy speech samples, and the difficult in determining the underlying affective state. The availability of podcasts containing long-time two-partners interactions by phone and the use of robust techniques for acoustic analysis can minimise the first two problems, whereas the focus on expressiveness change detection, and not on emotion recognition, overrides the third problem.

Scherer [11]'s componential appraisal model is well suited for the study of affective states that emerge from long-time interactions by phone. A central tenet of the theory is the postulation that the information processing subsystems of the organism (cognitive, motivational, physiological, motor) continuously check external and internal stimulus input using a set of functionally-defined SECs (Stimulus Evaluation Checks). These SECs monitor the environmental stimuli for novelty, pleasantness, goal/need significance, coping potential, and norm/self compatibility [11]. The outcome of these SECs act as a syndrome, engaging all the organism's subsystems, which characterises emotion. This model was used to predict vocal affect expression, including the integration with other modalities such as facial expression [14]. In the kind of interaction by phone studied here, the occurrence of fast changes in the expressiveness of one partner (the participant) due to events created by a second partner (the presenter) fits the continuous nature of the SECs proposed by Scherer. The main reason is because the presenter's action causes a series of immediate reactions on the participant, including motor responses that affect or relate to the dynamics of articulatory gestures which produce speech.

In the following, the material and the methodology for analysing expressiveness from both speaker- and listener-sides, as well as a procedure for the detection of changes in inferred expression are presented.

2. Measuring speech expressiveness

A set of 14 podcasts containing the record of 3 to 13 minutes of the interactions between two radio show presenters and 17 anonymous participants was used as scenarios for our analyses (in three podcasts two people talked to the presenter). The radio presenters of the show Programa do Chupim, from Rádio Metropolitana de São Paulo (http://metropolitanafm.uol.com.br/), Brazil, make prank phone calls either to people known to the presenters only, or to anonymous people offering a service in the advertisement section of a newspaper. All podcasts were downloaded on March 2008, and all participants signed a document accepting that the Radio put their dialogues on the site for public download. The 14 scenarios proposed by the presenters were considered real by all 17 participants, which yielded both positive and negative affective responses on the majority of the participants.

One of the advantages of this kind of interaction is the possibility of obtaining high levels of arousal of the affective responses due to the critical events introduced by the presenters. High levels of activation are not often found in the case of elicited emotions from control conditions [12, p. 232]. For each dialogue, emotion can be distinguished from other affective states by listening to the participants and applying Scherer's criteria [13], such as high intensity of response, perfect synchronisation with the antecedent, and high appraisal elicitation. However, in other to avoid the difficult in recognising the particular emotion in each case, as well as to allow equal treatment for all participants' utterances, no effort was made to directly characterise particular affective states. Thus, all participants' utterances were considered instances of expressive speech. Furthermore, to avoid the idiosyncratic use of the lexical items related to affect when judging expression, all utterances were evaluated using four expressive dimensions: the three classic emotional primitives of activation (relaxed-agitated), valence (unpleasant-pleasant), and dominance (controlled-under control) [10, 15], in addition to involvement (not involved-involved).

The reason for using these expressive dimensions is twofold: their levels can be more directly interpreted as outcomes of the SECs, and the use of the emotional primitives seems to produce higher levels of classification rate for emo-

tion recognition than the use of emotional lexical items [9].

2.1. Appraisal, acoustic analysis, and appraisal inference of expressive speech

In order to submit the 14 interactions for evaluation, 206 utterances were extracted from them. These utterances constitute a corpus containing the speech of the participants only, since the expression of the presenters lacked authenticity. Each utterance was delimited by the interruptions or turns of the respective presenter, lasting from 1.1 to 10.5 seconds.

An experiment was run out, which consists of two parts: the evaluation of the expressive space by a set of judges, and the prediction of the evaluation rates from a set of acoustic parameters. For predicting the rates, the set of utterances was split into two subsets, the training subset with 130 ramdomly-chosen utterances from 12 scenarios (scenarios 13 and 14 were put aside to be tested), and the test subset with the remaining 76 utterances (from the 14 scenarios, including 12 identical participants and two distinct, not present in the training subset). The training subset was evaluated by 12 judges, all of them undergraduate students of the first year in Linguistics. The evaluation was prepared as a MFC experiment in a Praat script [2]. Each judge received a sheet with the instructions for evaluating the respective expressive dimension. The four dimensions were evaluated by all judges in different days in two weeks. The sheet instructed the judge to choose one level from a 7-level semantic differential scale [10], by pointing with the mouse in the case corresponding better to what s/he just listened to, in terms of utterance's expression. For the activation dimension, for instance, the two poles of the scale were relaxed/calm (left) and agitated/stimulated (right). The case in the middle always indicated "neither one nor the other". All instructions were given in Portuguese. The task was accomplished before a computer screen by running the script under Praat. Each utterance was reproduced and a pure tone (1000 Hz) informed its end. After listening to the utterance, the judge pointed to his/her choice, and the script automatically reproduced the next utterance for the next choice until the end. The order of the utterances was ramdomised between judges. The whole task was completed in less than 25 minutes by all judges. The 7-level scale was linearly transformed to a scale between -1 and 1 (0 is the neutral choice) to ensure comparable magnitudes with the z-scored acoustic parameters. In order to both reduce the dimensionality of the expressive space, and avoid correlated dimensions, a Principal Component Analysis (using covariances) was applied to the original space. As shown below, two factors explained more than 95 % of the variance of the data.

To infer the judges' evaluation median rates for all utterances and dimensions, five classes of acoustic parameters were extracted: fundamental frequency (f_0), fundamental frequency first derivative (df_0), intensity, spectral tilt (SpTt), and Long-Term Average Spectrum (LTAS). See also [14]. One to four statistical descriptors were used for each class, producing twelve acoustic parameters: f_0 median, inter-quartile semi-amplitude, skewness, and 0.995 quantile; df_0 mean, standard-deviation, and skewness; intensity skewness; spectral tilt mean, standard-deviation, and skewness; and LTAS frequency standard-deviation. Spectral tilt is a correlate of vocal effort and was set to the difference of intensity in dB between the bands $0-1250$ Hz and $1250-4000$ Hz. The f_0 first derivative was used to detect abrupt changes in the melodic contour. The values of the f_0 and df_0 classes were z-scored by using f_0 mean and standard-deviation reference values for adult males: $(136, 58)$ Hz and females: $(231, 120)$ Hz. Spectral tilt was normalised by dividing its value by the complete-band intensity median, whereas LTAS standard-deviation was normalised by dividing its value by 10. Despite the clear presence of noise due to the telephone (which limits the signal energy between 300 and 3400 Hz), the Praat pitch auto-correlation algorithm is robust enough to avoid errors. Because spectral tilt is computed as a difference between two bands, the assumption of white noise minimises its effect on the parameter value. The same argument goes to the computation of the intensity skewness. Assuming white noise also ensure that the main source of variation for LTAS would be given by the signal itself. A script was implemented in Praat to automatically compute all these parameters.

A multiple regression analysis was used to associate the 12 acoustic parameters for each utterance with the two PCA-factored dimensions. Both techniques were applied by using the Statistica package, v. 8.0. The predicted and observed PCA-factored rates were compared with each other for the training subset only, because the utterances of the test subset were not evaluated by the judges.

By listening to the whole interactions, expressiveness seems to change according to changes in the situations proposed by the presenters of the radio show. That is why it is important to evaluate whether the judges are sensitive to changes in expressiveness in the case of no direct access to the chronological order of the events. Utterances were, thus, presented in random order across- and intra-speakers (note that they have no access to the presenter's turns). The analysis of the utterances of specific participants is, however, done chronologically in order to point out the dimensions and/or acoustic parameters that signal expressiveness change.

3. Results

Inter-rater reliability was computed by obtaining the kappa index (k). In order to do so, the 7-level transformed scale was divided into three categories: inferior (-1 and $-0,67$), middle -0.33 to 0.33), and superior (0.67 and 1). Significance for $\alpha = 0.001$ is achieved with $z > 3.09$. Fair inter-rater reliability was obtained for activation ($k = 0.38, z = 47.3$), valence ($k = 0.37, z = 45.3$), and involvement ($k = 0.26, z = 28.6$). Though significant ($k = 0.05, z = 6.4$), the dominance dimension was discarded from the analysis due to the low inter-rater reliability. See [1] for similar figures for kappa. The levels of the three other dimensions scatter from -1 to 1 with means of 0.20 (activation), -0.21 (valence), and 0.27 (involvement). The lowest absolute value of the correlation coefficient for any two dimensions is 0.80.

The PCA analysis revealed that two factors explain 97 % of the variance. Factor 1 explains almost 90 % of the variance and is an even combination of the three dimensions: Fact 1 = 0.59(Act-Act$_{mean}$) - 0.59(Val-Val$_{mean}$) + 0.55(Inv-Inv$_{mean}$). Factor 2 explains around 7 % of the variance: Fact 2 = 0.17(Act-Act$_{mean}$) + 0.73(Val-Val$_{mean}$) + 0.63(Inv-Inv$_{mean}$). After listening carefully to the utterances we decided to call factor 1, promptness, and factor 2, empathy. The latter is related to the interaction between partners by means of turn-exchange cues. The PCA-transformed expressive space can be seen in Fig 1. An ANOVA with SCENARIO as independent variable reveals a difference among the 12 scenarios of the training subset for the two PCA factors (only one participant per scenario was considered for ANOVA). For promptness, $F(11, 105) = 51.7, p < 10^{-7}$ with 7 distinct groups (unequal N HSD post-hoc analysis for homogeneous groups); for empathy, $F(11, 105) = 4.2, p < 10^{-4}$

with 2 distinct groups. Scenarios 4 and 5 form the group with the highest levels of promptness. Scenario 5 is the situation where the father (the participant) thinks her daughter is pregnant. Scenarios 2 and 10 are the less empathic for which the participants seem relatively indifferent to the situations proposed by the presenter.

The multiple regression analysis applied to the training subset revealed the relevance of two classes of parameters: df_0 (mean), and spectral tilt (mean and standard deviation) for predicting the levels of promptness ($R = 67$ %, training subset). In fact, an ANOVA of spectral tilt mean as dependent variable reveals that this parameter separates 7 different groups with $F(11, 105) = 123.5, p < 10^{-8}$. The participant of scenario 10 has the lowest level of vocal effort (higher values of spectral tilt), and the participants of scenarios 5 and 6, the highest. Equation 1 shows the regression for promptness: df_0 mean, related to f_0 change, and Spectral Tilt mean and standard-deviation, related to vocal effort, are the only relevant parameters to determine promptness: the higher the vocal effort (negative value of spectral tilt) and its variability, the higher the level of promptness.

$$promptn. = 0.29 - 0.17mean_{df_0} - 14.05mean_{SpTt} + \\ +10.48SD_{SpTt} \quad (1)$$

Empathy was predicted with four classes of parameters, as can be seen in equation 2 ($R = 40$ %, training subset). The increase of vocal effort and spectral variabilities decrease empathy, whereas f_0 high levels (represented by the 0.995 quantile) increase empathy. These features are related to a relaxed exchange between partners.

$$emp. = 0.360.12medf_0 + 0.08f_0(99.5quant) - \\ -0.48f_0(skew) - 0.05df_0(skew) + 4.64SpecTilt(mean) - \\ 4.81SpecTilt(SD)0.12LTAS(SD) \quad (2)$$

The predictions for the training set can be seen in Fig. 1. As

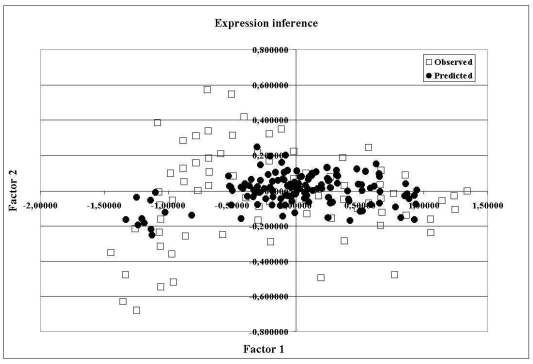

Figure 1: *Predicted and observed expressive space for the training subset.*

expected from the variance explained by each factor, the scattering of factor 1 (promptness) is much more covered than the one of factor 2 (empathy). By using three categories of rates, errors of changing the inferior category for the superior category and vice-versa were verified only for promptness: 7 out 130 utterances of the training set, and 7 out 76 utterances of the test set. The expression inference via regression from the acoustic parameters automatically extracted from the utterance was then used to study expressiveness change in all 14 scenarios.

3.1. Changes in expressiveness in two domains

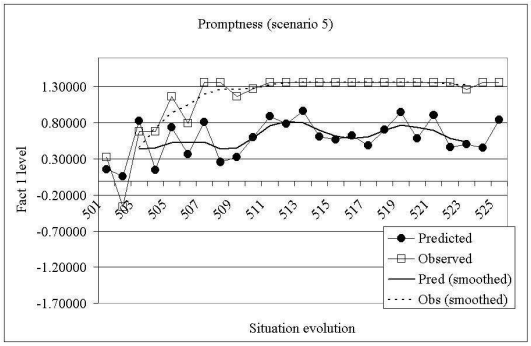

Figure 2: *Predicted and observed promptness for scenario 5.*

Fig. 2 shows the observed and predicted values of promptness level for the participant of scenario 5, a father who thinks his daughter is pregnant. A classic 5-point moving average was applied to obtain smoothed versions of both contours. From utterances 1 to 5 the participant talks to his daughter, and from utterance 6 on, to the radio presenter. The observed contour shows, as evaluated by the judges, a saturation to a maximum level of promptness. This is not the case of the predicted contour, which shows a trend to higher levels of promptness with oscillations. The predicted levels are entirely based on acoustic parameters, contrary to the observed rates. These latter are also dependent on other influences, such as the semantic weight of the lexical items. In this situation, it is likely that the judges inferred the reasons for the participant's rage and decided to choose maximum levels of activation and involvement, and minimum valence, given the lexical items used by the participant. On the other hand, any person unable to understand Portuguese can detect a drop in promptness (to action) by listening to the audio files of utterances 7 to 8 (sc507.wav, and sc508.wav), as signalled by the predicted contour. No change is signalled in the observed contour, however.

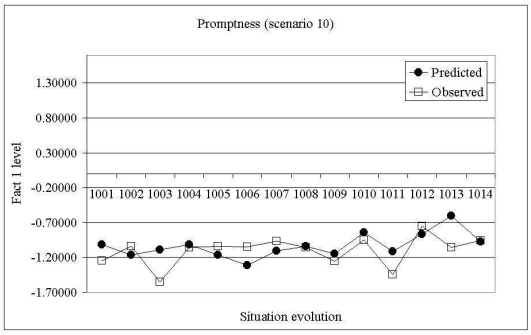

Figure 3: *Predicted and observed promptness for scenario 10.*

Fig. 3 shows the observed and predicted promptness contours for the participant of scenario 10, a guy selling a pirate CD. It can be seen that there is almost no change in the rather low level of promptness, even when the participant respond to the identification of two presenters as policemen (utterance 10

on). However, a slight increase of both predicted and observed promptness can be seen from that moment on.

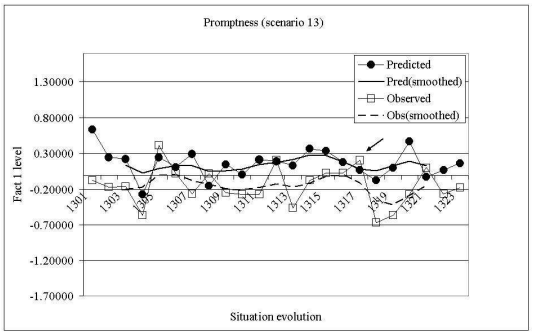

Figure 4: *Predicted and observed promptness for scenario 13.*

Fig. 4 shows the observed and predicted promptness contours for the participant of scenario 13, a girl thinking to talk to a cousin she did not see for a long time. Her reactions are very positive. From utterance 4 to 5 she thinks to have discovered the identity of this cousin, and her level of promptness increases (files sc1304.wav, and sc1305.wav). There is a drop in level in both the predicted and observed contours from utterance 17 to 18 (files sc1317.wav, and sc1318.wav), as indicated by the arrow. Utterance 18 is her reply to a problem happened with her actual aunt (the presenter's fake mother).

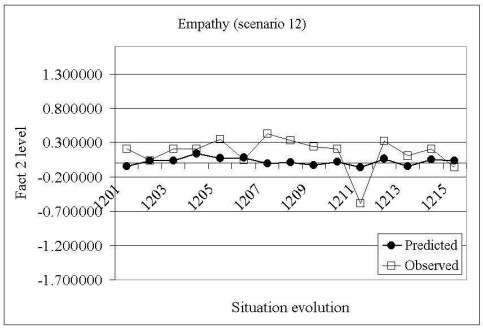

Figure 5: *Predicted and observed empathy for scenario 12.*

Fig. 5 shows the observed and predicted empathy contours for the participant of scenario 12, a woman offering a cattering service for a birthday part. Her reactions are relatively positive especially at the end. The levels of empathy are not so well predicted as expected by the lower regression coefficient, and the lower variance explanatory power from the PCA. Predicted levels increases seem to correspond to observed increases. The drop in empathy from utterance 10 to 11, correspond to a change from asking the kind of food for the party, to writing down and saying it aloud at the same time.

4. Summary and perspective

Static values for and changes in expressiveness can be reliably identified from acoustic-parameters-derived levels of promptness. An extensive training on corpora containing other situations could ensure the representative of others dimensions, including empathy, which is more related to pleasantness and involvement. The procedure shown here seems promising for the automatic detection of expressiveness change, depending on the amount of variance explained by each factor. The procedure can also be used to relate content with expression by applying the same prediction equations to previously automatically or manually delimited discourse [4] or prosodic constituents. This is useful in automatic recognition systems for detecting changes in human behaviour during interaction with a machine.

5. Acknowledgments

This work was sponsored by Natura Tecnologia e Inovação de Produtos Ltda. The work enriched after discussions with C. Albuquerque, F. Mello, and C. Pellegrino, and suggestions by S. Madureira. The students L. Lucente and T. Vale assisted during the evaluation tests. A grant from the CNPq (300296/2005-3) is also acknowledged.

6. References

[1] Alm, C. O., Sproat, R., "Perceptions of emotions in expressive storytelling", Proc. Interspeech, Lisbon, 2005.

[2] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" (Version 5.0.35) [Computer program], Online: http://www.praat.org, accessed in 2008.

[3] Grimm, M., Kroschel, K., Narayanan, S., "Modeling emotion expression and perception behavior in auditive emotion evaluation", Proc. Speech Prosody 2006 [CD], Dresden, 2006.

[4] Grosz, B. J., Sidner, C. L., "Attention, intention, and the structure of discourse", Comp. Ling., 12:175–204, 1986.

[5] Ito, K., Speer, S. R., "Using interactive tasks to elicit natural dialogue", in Sudhoff, S. et al. [Ed], Methods in empirical prosody research, 229–257, Walter de Gruyter, 2006.

[6] Johnstone, T., "Emotional speech elicited using computer games", Proc. 4th ICSLP [CD], Philadelphia, 3, 1996.

[7] Kehrein, R., "The prosody of authentic emotions", Proc. Speech Prosody 2002 [CD], Aix-en-Provence, 2002.

[8] Kohler, K., "Paradigms in experimental prosodic analysis: from measurement to function", in Sudhoff, S. et al. [Ed], Methods in empirical prosody research, 123–152, Walter de Gruyter, 2006.

[9] Lugger, M., Yang, B., "An incremental analysis of different feature groups in speaker independent emotion recognition", Proc. XVIth ICPhS, Saarbrücken, 2149–2152, 2007.

[10] Osgood, C. E., Suci, G. J., Tannenbaum, P. H., The measurement of meaning, Chicago; Urbana: UIP, 1957.

[11] Scherer, K. R., "Vocal affect expression: a review and a model for future research", Psychological Bulletin, 99:143–165, 1986.

[12] Scherer, K. R., "Vocal communication of emotion: a review of research paradigms", Speech Communication, 40:227–256, 2003.

[13] Scherer, K. R., Bänziger, T., "Emotional expression in prosody: a review and an agenda for future research", Proc. Speech Prosody 2004, Nara, CD Proc, 2004.

[14] Scherer, K. R., Ellgring, H., "Multimodal expression of emotion: affect programs or componential appraisal patterns?", Emotion, 7 (1):158–171, 2007.

[15] Schlosberg, H., "Three dimensions of emotion", The Psychological Review, 61 (2):81–88, 1954.