Vocal Gestures in Slovak: Emotions and Prosody

Štefan Beňuš[1,2] and Milan Rusko[2]

[1] Constantine the Philosopher University, Stefanikova 67, 94974 Nitra, Slovakia
[2] Institute of Informatics of the Slovak Academy of Sciences, Dubravska cesta 9, 845 07 Bratislava, Slovakia
sbenus@ukf.sk, milan.rusko@savba.sk

Abstract. Hot-spot words are indicators of high emotional invovement of speakers in the conversation and contain cues to the emotional state of the speaker. Understanding and modeling of these cues may improve the effectiveness and naturalness of automated cross-modal dialogue systems. In this paper we investigate the relationship between prosody and emotions in a subgroup of hot-spot words: non-verbal vocal gestures with a problematic textual representation. We extracted these gestures from a recording of a puppet play and argue that this corpus is well suited for investigating emotional speech. We identify the expressive load of non-verbal vocal gestures in Slovak and report on multiple ambiguities in their emotional and discourse functions. The relationship between prosody and emotions in non-verbal hot-spot words is very complex and a ToBI-based framework of discrete representation of prosody is useful but not sufficient for modeling this relationship.

Keywords: Nonverbal vocal gestures, emotional speech, prosody, ToBI.

1 Introduction

Signals about the internal state of the speaker are both visual and auditory. For example deception could be cued by gestures such as putting a hand over the mouth [1] but it may also correlate with the use of filled pauses [2] or higher pitch [3]. Similar examples of cross-modal characteristics could be found for many types of emotions. Hence, a comprehensive model of human communication relies on both visual and auditory characteristics.

Efforts in understanding and modeling the visual and auditory aspects of the internal state of a speaker are essential in improving applications relying on human-machine communication. For example, in automated call-center applications, detecting a frustrated or angry customer is important so that s/he can be switched to a human operator. Or, the efficiency of tutoring systems is increased if they can reliably detect uncertainty. Generation of good quality emotional dialogues is usable in gaming and other industries.

One of the promising approaches in utilizing emotions in speech is related to so-called hot-spot words. The presence and acoustic characteristic of these words may facilitate the detection of high involvement in meetings and thus point to important information for automatic retrieval [4]. Additionally, prosody of these ambiguous hot-spot words can signal emotional valence of the speaker, that is if the speaker's

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 223–231, 2009.

emotions are positive or negative (e.g. [5] on *'whatever'*). Therefore, as a long-term goal, the presence and the acoustic and prosodic characteristics of a hot-spot word may be used for measuring the expressiveness of larger units of analysis such as utterances, speaker turns, or whole conversations.

In this paper we concentrate on the auditory expressions of emotion of special type of hot-spot words that we call non-verbal vocal gestures. (see [6] for the taxonomy of gestures in the visual mode). Following [6] we consider non-verbal speech gestures those "speech sounds that have their own phonetic content, phonological function, and prosody, but [they] do not have an adequate linguistic (or text) representation". In this view, non-verbal gestures belong among hot-spot words with significant expressive power and information about the speaker's internal state. The major question is how to describe the information they contain so that it could be utilized to improve the effectiveness of systems based on the recognition and generation of emotional speech. In the following, we discuss a scheme for labeling the prosody of these gestures and report on an experiment testing the perception of emotions signaled by these gestures.

2 Prosody Description of Emotional Speech

The best way of studying how prosody conveys emotions and attitudes is when prosodic features are represented in a theoretical framework of a prosodic model [7]. Previous research has shown that both discrete and continuous features that capture acoustic and prosodic information and the voice quality features are useful in characterizing emotional speech. For example, [8] supported the view that automatically extracted pitch and energy features correlate with the activation of emotions; that is to what degree the speaker is emotionally involved. Interestingly, discrete features that capture the intonational contour were useful in characterizing the valence of emotions, that is, if the emotion is positive or negative. More specifically, [8] used the ToBI system of intonation description [9] and in their corpus, utterances with plateau contours (H-L%) were positively correlated with negative emotions while utterances with falling contours (L-L%) correlated with positive emotions.

The use of ToBI for describing the prosodic characteristics of emotional speech was also supported by proponents of the British school of intonation (e.g. [10]). Hence, given the usefulness of the ToBI framework for capturing important information in emotional speech, we decided to continue our efforts at building a ToBI version suitable for describing Slovak intonation (henceforth SK-ToBI, [11]) and test its usefulness for describing non-vocal gestures.

2.1 ToBI System

ToBI [9] is a model that captures phonologically and pragmatically meaningful aspects of intonation with the use of a finite set of discrete symbols. Primarily, ToBI is concerned with identifying intonationally prominent events as pitch targets (=Tones) and the degree of disjuncture between pairs of words (=Break Indexes). ToBI labeling scheme was developed by linguists and speech scientists in order to study the functional differences conveyed by intonation and to enable sharing speech corpora labeled in a uniform scheme due to the fact that labeling intonation is very time and effort consuming.

The original ToBI developed for Mainstream American English (MAE) recognizes three tonal targets: H(igh), L(ow), and !H that represents a down stepped high target after another high target within a single intonational phrase. There are three types of tonal targets: pitch accents that signal intonational prominence on a given word, marked with '*', and boundary tones that signal the right edge of the intermediate and intonational units marked with '-' and '%' respectively. This system then gives the following inventory:

- Pitch accents: H*, !H*, L*, L+H*, L+!H*, L*+H, L*+!H, H+!H*
- Phrase accents: H-, !H-, L-
- Boundary tones: H%, L%

Every intonational boundary is assumed to contain both phrase accent and boundary tone while intonational phrase is only marked with a phrase accent. The ToBI framework has been adapted for the description of intonation of many unrelated languages (see [12] for a review).

2.2 SK-ToBI

Slovak is a west-Slavic language spoken by roughly 5 million people in Slovakia. Our current ToBI framework for description of Slovak intonation is based on a corpus of 246 utterances that could be divided into two groups. The first group is a set of 74 sentences balanced for syntactic and pragmatic meaning, and read by a single speaker. The second part of the corpus comes from a recording of a play acted by a single puppet-player [14]. We selected the utterances of 2 characters (Faust and Jester) with similar lengths of the speech material (79 and 93 utterances respectively).

All material was transcribed, manually aligned with the speech signal, and labeled following the annotation conventions developed for MAE ToBI. Using PRAAT [13], three annotators participated in labeling; one is an expert annotator of MAE ToBI, the other two are beginners. Inter-labeler agreement was measured by Fleiss' κ [15]. The agreement for presence vs. absence of pitch accent and boundary tones was measured at 0.74 and 0.76 respectively, where values between 0.6 and 0.8 correspond to substantial agreement. The agreements for the type of the pitch accent and the boundary tone were worse (below 0.5).

The analyses of our labeled corpus showed that ToBI is a suitable system for capturing the range of differences cued by intonation in Slovak. We identified three areas where SK-ToBI might differ from MAE ToBI but more research is needed to see if the changes of the inventory of tones are warranted or just the alignment of tonal targets with the segmental material is different.

First, H-L% boundary tone in MAE ToBI describes so called plateau intonation, especially after a preceding H* pitch accent. In Slovak, a similar plateau intonation is possible and H-L% was used to label it. However, yes/no questions in Slovak tend to rise toward and during the final pitch accented syllable but then the pitch stays flat on the following non-stressed syllables. Current MAE ToBI does not give other option but to use H-L% for this type of contour. Hence H-L% in Slovak is ambiguous between a question rise and a plateau, both of which occur frequently, and some utterances with identical H-L% labeling in the two languages sound significantly different.

The second, issue is the presence of trailing tones that seem to mark the right edge of a prosodic word in Slovak. A similar proposal for tones that mark the edges of words has been made for Catalan [16] but no generally accepted addition to the ToBI framework has been made.

Finally, Slovak may require the addition of H*+L tone, already suggested for German [17] and other languages while contrastive function of some rare tones such as L*+!H or !H-H% is questionable in Slovak.

3 Corpus of Non-verbal Vocal Gestures

Designing a corpus suitable for experimental investigation of emotional speech represents significant obstacles. Spontaneous unscripted speech of good sound quality rarely contains a rich variety of emotions. Hence, the primary method of data acquisition in the past was to ask professional actors to act emotions on semantically neutral utterances such as series of numbers. This approach allows for a rich source of emotions and control of other variables such as linguistic context. However, elicited acted emotions are not realized identically to the ones occurring in spontaneous speech. More specifically, the elicited emotions are more exaggerated than the real ones and many times may sound artificial and stilted. On the other hand, [18] showed that although acted emotions are more exaggerated than real-life ones, the relationship of these two types of emotions to acoustic correlates is not contradictory.

In our approach, we use the corpus of a play acted by a single puppet-player [13]. The corpus is rich with emotional speech since the visual cues to emotions are significantly limited for characters in a puppet play. Hence, the speech of the puppet player must be expressive. On the other hand, despite the existence of a script, the emotions are not really elicited, but they are expressed semi-spontaneously as it is required by the plot and the development of the play. Our corpus of puppet plays thus represents an intermediate level between elicited acted speech of actors and spontaneous everyday communication. Therefore, we believe that our corpus has potential to improve our understanding of the system that underlies the production and perception of features linked to the emotional state of the speaker.

Table 1. Frequency distribution of non-verbal vocal gestures in the corpus. The transcriptions in the square brackets are approximated broad IPA transcriptions.

NVG	N	%	NVG	N	%
[joj]	15	16.9	[ejha]	3	3.4
[he]	12	13.5	[jej]	3	3.4
[aha]	10	11.2	cry	2	2.2
[e]	8	9	[heja]	2	2.2
[jaj]	8	9	[hm]	2	2.2
laugh	8	9	[mmhm]	2	2.2
inhale	4	4.5	[e-e]	1	1.1
[juj]	4	4.5	exhale	1	1.1
[a]	3	3.4	[mm]	1	1.1

In order to study the form and function of non-verbal gestures we extracted 89 instances of non-verbal vocal gestures from the transcribed and aligned recording of the puppet play. We extracted the instances of both true non-verbal gestures such as laughs, cries, inhale and exhale sounds, or filled pauses, as well as semi-verbal gestures that represent particles such as [joj], [juj], [he], [a], [e], or [aha], and could convey a range of emotions. The distribution of the non-verbal gestures is summarized in Table 1.

The prosody of these gestures was transcribed using SK-ToBI. It should be noted that the labeling of true non-verbal gestures such as inhales, laughs or grunts was very problematic in SK-ToBI due to the fact that the activity of the vocal folds (and consequently the F0 contour) is often very limited. The labeling of semi-verbal gestures was much more straightforward. Here we concentrate on the Tones labeling since each non-verbal gesture constituted a prosodic phrase on its own and thus the juncture following each gesture was identical in terms of ToBI.

In terms of the number of pitch accents, all but two tokens received a single accent. The two non-verbal gestures with double accents were multisyllabic: [jojojojojoj] and a laugh, both of them had high accent o the initial syllable and a low(er) accent on the final syllable of the gesture. The distribution of the types of pitch accent and boundary tone in our corpus are shown in Table 2 below, and the distribution of complete intonational contours is in Table 3.

Table 2. Frequency distribution of the pitch accents and boundary tones among the non-verbal vocal gestures

Pitch accent			**Boundary tone**		
Type	**N**	**%**	**Type**	**N**	**%**
H*	54	60.7	L-L%	39	43.8
L+H*	16	18.0	H-L%	27	30.3
L*	13	14.6	!H-L%	14	15.7
H+{!H*, L*}	6	6.7	{H, !H, L}-H%	9	10.1

Table 3. Frequency distribution of the intonational contours labeled with the ToBI scheme among the non-verbal vocal gestures

Intonational contours					
Type	**N**	**%**	**Type**	**N**	**%**
H*L-L%	20	22.5	L+H*H-L%	9	11
H*H-L%	16	18	L+H*L-L%	5	5.6
H*!H-L%	14	15.7	H*{H, !H, L}-H%	5	5.6
L*L-L%	11	12.4	H+!H*L-L%	3	3.4
			Other	6	6.6

4 Prosody and Emotions of Non-verbal Gestures

To study the relationship between the form of the non-verbal gestures and the emotions they conveyed, we ran a pilot perception experiment. From the set of six basic

emotions detectable from facial behavior [1] we selected five emotions: ANGER, JOY, SADNESS, FEAR, and SURPRISE and asked 6 subjects to judge the presence of these emotions in non-verbal gestures on the scale from 0 (no emotion) to 3 (very strong emotion). Subjects listened to the audio signal of the non-verbal gesture alone, without any surrounding context or transcription. Each rating was normalized by participant using z-scores normalization in order to account for the variation the subjects might have produced in the use of the scale. We then calculated the mean and standard deviation values for each token and emotion.

4.1 Results and Discussion

As a general measure of expressive load we took the sum of the mean z-score values for each token and emotion (the higher the value, the greater the expressive load). We also calculated a measure of agreement among the subjects, which we took as a sum of the standard deviations for each token and emotion (the higher the value, the worse the agreement). We found a highly significant positive correlation between these two general measures in our data; $r(1, 89) = 0.699$, $p < 0.01$. That is, our subjects agreed better on less expressive tokens than on the more expressive ones. Hence, the identification of the lack of emotions just from the acoustics of non-verbal gestures is easier than the identification of emotions.

Gestures with low expressive load and high annotators' agreement include backchannels [mmhm] and particles [he]/[e], and [aha]. 15 of the 20 least expressive gestures belong to this group. The primary function of backchannel [mmhm], is to acknowledge some previous event (either verbal or non-verbal). Particle [he] also seems to have rather discursive than emotive function: it was mostly used to mark the end of a speaker's turn. Particle [e] was mostly used to signal a speech error, which is similar to one of the primary functions of filled pause. Finally, [aha] may have multiple functions, one of which is to acknowledge previous information; hence, it may function as a backchannel. It is interesting that all three [aha] tokens among the group of least expressive gestures were relatively short and labeled L+H*H-L%, which is consistent with the backchannel function. Other functions of [aha] include expressing surprise and drawing attention to something, and were produced in different prosody. In sum, there are several non-verbal gestures in our corpus that seem to function primarily in managing the flow of conversation, and it is thus reasonable that these don't have high expressive load and should not be considered hot-spot items.

At the other end of the continuum are vocal gestures with high expressive load but low annotators' agreement. The most common among these gestures are clearly [joj], [jaj], and [juj], and also two inhale sounds. All [jVj] gestures are very expressive in Slovak, can contain multiple syllables within one gesture, and are able to signal a broad range of emotions. In our corpus, they were associated mostly with fear and sadness, and then with surprise. A high pitch accent (H*) was common for these emotions, followed with either a fall (L-L%) or an unfinished fall ({H, !H}-L%).

Moving now to describing the relationship between intonational contours and emotions, Fig. 1 below represents the most frequent melody contours ($N > 2$) in our corpus as described by SK-ToBI and how they signaled particular emotions.

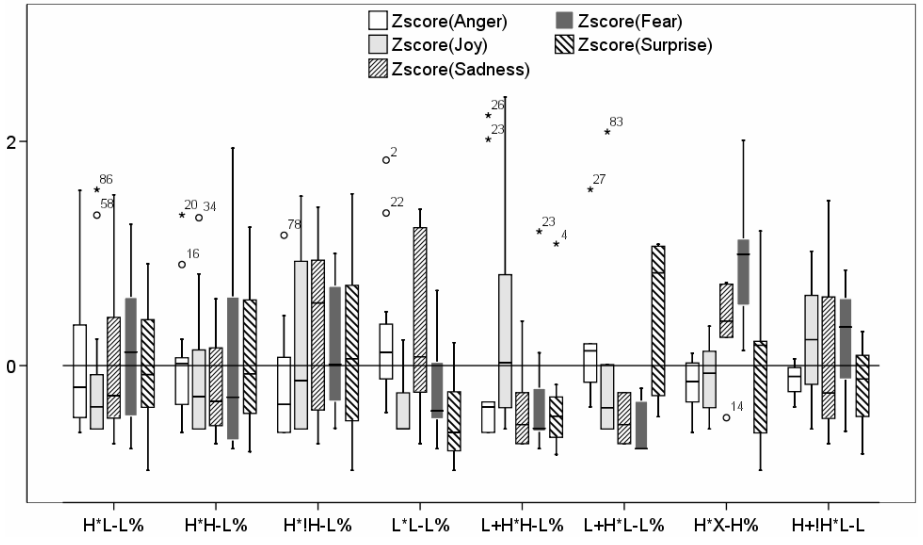

Fig. 1. Boxplots representing the intonational contours on the x-axis and mean z-score values for emotions on the y-axis. The contours are in descending ordered of frequency from the left. The bottom and the top of the box represent lower and upper quartiles of the distribution respectively, the band near the middle is the median, and the whiskers represent the minimum and the maximum values. Circles represent outlier values and asterisks extreme outlier values.

Despite the density of the data in Fig. 1, it provides two types of useful information. First, it identifies the outliers in the distributions. Outliers represent extreme emotions and are thus true hot-spot words of our corpus. The figure identifies 15 outliers, one of which is a 'negative outlier' and cues the absence rather than presence of emotion. Of the remaining 14 true hot-spot gestures, the most outliers were identified for ANGER (8), and JOY (4). Out of 8 outliers for ANGER there are 4 [he] tokens, 2 [ejha], [juj], and [a]. High anger ratings for some [he] tokens are surprising in the view of its discourse function described above. Hence, both [he] and [aha] are highly ambiguous in terms of whether they are emotionally loaded or not. Out of 4 outliers for JOY there are 2 laughs, [he] and [jaj].

The nature of an outlier in this Figure is that it has the same melody contour as other vocal gestures that don't cue a particular emotion but this outlier token cues extreme emotion. Because of this and due to the fact that there is no clear pattern in terms of the intonational contours for these outliers, we speculate that other features such as intensity, duration, pitch range or voice quality are more important in cuing these extreme emotions than the intonational contour itself.

The second information present in the Figure is that some tendencies in the relationship between the intonational contour and the other three emotions can be observed. For example, pitch accents with a clear low region (L* and L+H*) tend to signal the absence of FEAR while the presence of final rise (H%) signals its presence. However, more robust patterns are missing and further research is needed to decode the complex relationship between emotions and intonation contours.

5 Conclusion

In this paper we set to investigate the relationship between prosody and emotions in non-verbal vocal gestures extracted from a recording of a puppet play in Slovak. We reported that ToBI framework is suitable for describing Slovak intonational patterns and we discussed several issues where this framework for American English and Slovak might differ. We identified the expressive load of non-verbal vocal gestures and discovered multiple ambiguities in their emotional and discourse functions. The prosody-emotion relationship in non-verbal hot-spot words is very complex and a ToBI framework is useful but definitely not sufficient for modeling this relationship.

In future research we plan to investigate the influence of continuous features such as intensity, pitch range, duration, spectral tilt, jitter, or harmonics-to-noise ratio on the perception of emotional speech in these hot-spot words. We agree with one of the reviewers that surrounding context, namely intonation contour and other prosodic features before or after the hot-spot words, could also provide useful information. Currently we work on designing an experiment that would test the effect of lexical and prosodic context on the perceived emotions of non-verbal vocal gestures.

Acknowledgments. We would like to thank R. Kováč, R. Sabo, and the subjects in our perception experiments. This work was supported by the of the Ministry of Education of the Slovak Republic, Scientific Grant Agency project number 2/0138/08 Applied Research project number AV 4/0006/07, **by the Slovak Research and Development Agency under the contract No. APVV-0369-07,** and by the European Education, Audiovisual and Culture Executive Agency LLP project EURONOUNCE.

This project has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

1. Ekman, P., Friesen, W.V., Ellsworth, P.: What Emotion Categories or Dimensions can Observers Judge from Facial Behavior? In: Ekman, P. (ed.) Emotion in the Human Face, pp. 39–55. Cambridge University Press, Cambridge (1982)
2. Benus, S., Enos, F., Hirschberg, J., Shriberg, E.: Pauses and Deceptive Speech. In: Hoffmann, R., Mixdorff, H. (eds.) Proceedings of 3rd International Conference on Speech Prosody. TUD Press, Dresden (2006)
3. DePaulo, B.M., Lindsay, J., Malone, E., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to Deception. Psychological Bulletin 129(1), 74–118 (2003)

4. Wrede, B., Shriberg, E.: Spotting Hotspots in Meetings: Human Judgments and Prosodic Cues. In: Proceedings of European Conference on Speech Communication and Technology, pp. 2805–2808 (2003)
5. Benus, S., Gravano, A., Hirschberg, J.: Prosody, Emotions, and...'whatever'. In: Proceedings of International Conference on Speech Communication and Technology, pp. 2629–2632 (2007)
6. Rusko, M., Juhár, J.: Towards Annotation of Nonverbal Vocal Gestures in Slovak. In: Cross-Modal Analysis of Verbal and Nonverbal Communication. LNCS (in press)
7. Mozziconacci, S.: Prosody and Emotions. In: Bel, B., Marlien, I. (eds.) Proceedings of 1st International Conference on Speech Prosody, pp. 1–9 (2002)
8. Liscombe, J., Venditti, J., Hirschberg, J.: Classifying Subject Ratings of Emotional Speech Using Acoustic Features. In: Proceedings of European Conference on Speech Communication and Technology, pp. 725–728 (2003)
9. Beckman, M.E., Hirschberg, J., Shattuck-Hufnagel, S.: The Original ToBI System and the Evolution of the ToBI Framework. In: Jun, S.-A. (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 9–54. Oxford University Press, Oxford (2005)
10. Roach, P.: Techniques for the Phonetic Description of Emotional Speech. In: Proceedings of Speech Emotion 2000, pp. 53–59 (2000)
11. Rusko, M., Sabo, R., Dzúr, M.: Sk-ToBI Scheme for Phonological Prosody Annotation in Slovak. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 334–341. Springer, Heidelberg (2007)
12. Jun, S.-A.: Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, Oxford (2005)
13. Rusko, M., Hamar, J.: Character Identity Expression in Vocal Performance of Traditional Puppeteers. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 509–516. Springer, Heidelberg (2006)
14. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer, `http://www.praat.org`
15. Fleiss, J.L.: Measuring Nominal Scale Agreement among Many Raters. Psychological Bulletin 76(5), 378–382 (1971)
16. Estebas-Vilaplana, E.: Catalan Pre-Nuclear Accents: Evidence for Word-Edge Tones. In: Solé, M.J., Recasens, D., Romero, J. (eds.) Proceedings of the 16th International Congress of Phonetic Sciences, pp. 1779–1782 (2003)
17. Grice, M., Baumann, S., Benzmüller, R.: German Intonation in Autosegmental-Metrical Phonology. In: Jun, S.-A. (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 55–83. Oxford University Press, Oxford (2005)
18. Williams, C.E., Stevens, K.N.: Emotions and speech: Some acoustical factors. Journal of the Acoustical Society of America 52, 1238–1250 (1972)