Towards Annotation of Nonverbal Vocal Gestures in Slovak

Milan Rusko[1], Jozef Juhár[2]

[1] Institute of Informatics of the Slovak Academy of Sciences, Dubravska cesta 9, Bratislava,
[2]Faculty of Electrotechnical Engineering, Letna 9, Technical University Košice,
Slovak Republic
milan.rusko@savba.sk, jozef.juhar@fei.tuke.sk

Abstract. The paper presents some of the problems of classification and annotation of speech sounds that have their own phonetic content, phonological function, and prosody, but they do not have an adequate linguistic (or text) representation. One of the most important facts about these "nonverbal vocal gestures" is that they often have a rich semantic content and they play an important role in expressive speech. The techniques that have been used in an effort to find an adequate classification system and annotation scheme for these gestures include prosody modeling and approaches comparing the nonverbal vocal gestures with their verbal (lexical) and body counterparts.

Keywords: Nonverbal vocal gestures, non lexical speech sounds, paralinguistic speech, grunts.

Introduction

Building of speech databases including TV debates recording, court proceedings, and dialogues have started in Slovakia recently. These databases will become parts of the Slovak National Corpus [1] that has only had a text part up till now. For a deeper study of expressive speech phenomena a sophisticated classification and description scheme of non-lexical speech sounds in Slovak is needed. This paper presents ideas and concepts that will hopefully give rise to a first version of such a classification scheme. As there was no expressive speech database available, the speech corpus based approach was not possible. Therefore we decided to take the advantage of existence of the Japanese-Chinese-American-Slovak Picture dictionary of gestures [2]. We tried to find out which of these gestures have their vocal counterpart or accompanying sound. This analysis made it possible to define first set of candidates for our list of “nonverbal vocal gestures”. This was further enriched by candidates obtained from Slovak National Corpus and from tools for crossword solvers and scrabble players. Problems of description - orthographic representation, acoustics/phonetics, prosody, semantics and pragmatics of this special class of speech sounds are presented in this paper.

Nonverbal Information in the Speech Signal

The information communicated in spoken language can be categorized as linguistic, paralinguistic, and extra-linguistic [3]. The verbal content, the actual meaning of the words, is thought of as linguistic information. The extra-linguistic channel contains information about the speaker's basic state and culture. The paralinguistic channel carries information about momentary changes in the usual (extra-linguistic) baseline, such as expression of emotions etc.

In this study we will aim at nonverbal vocal gestures (NVG) that we define as speech sounds that have their own phonetic content, phonological function, and prosody, but they do not have an adequate linguistic (or text) representation. One of the most important facts about these non-lexical speech sounds [4] is that they often have a rich semantic content and they play an important role in expressive speech.

Method

In our work we want to utilize the knowledge and methods used in verbal communication research: linguistics, orthography, orthoepy, phonetics, phonology, speech acoustics, pragmatics, nonverbal non-vocal gestures research, and know-how coming from speech synthesis and speech recognition research. The first draft annotation scheme will be used for test annotations and will then be modified according to their results. In the second round a bigger database will be annotated and an inter-annotator agreement will be tested. The resulting detailed annotation should enable clustering the whole space of nonverbal speech gestures into subspaces (groups) according to various features (e.g. similar acoustic structure, semantics etc.).

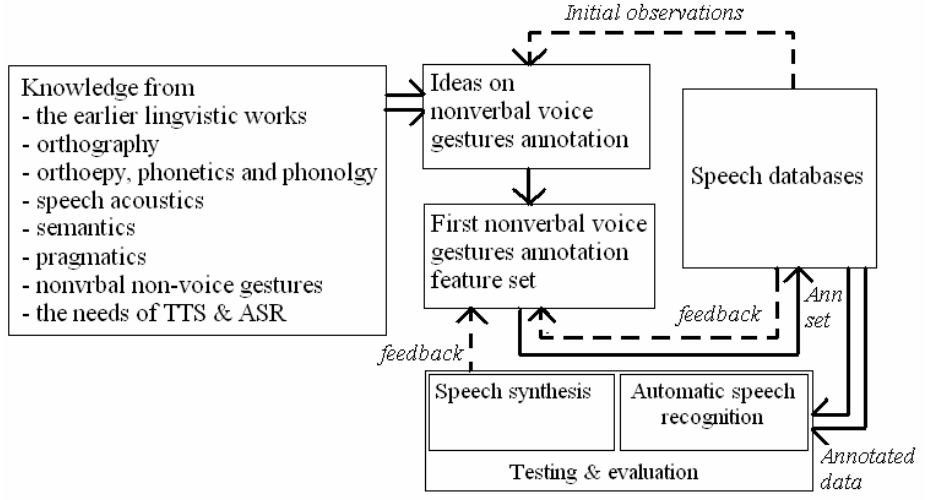

Fig. 1. Schematic diagram of our approach to nonverbal speech gestures classification scheme design and testing.

The definitive version of the annotation scheme will be used to annotate training speech database and speech synthesis database and tested in automatic speech recognition and unit-selection speech synthesis.

Gestures

A gesture is a form of non-verbal communication made with a part of the body, used instead of or in combination with verbal communication [5].

As we did not have any speech material, nor classification or description conventions available to start with in our research, we decided to take the advantage of the existence of the Japanese-Chinese-American-Slovak Picture dictionary of gestures [2]. The material of the dictionary comprises four groups according to the generic meaning. Such division is directed by semantics. According to the author the gestures express physical body, initiative contact, emotional body or mental body. The optics of the dictionary is directed towards only one body at a time; both verbal and non-verbal communication is taken into account.

We checked the list (classification, pictures) of body gestures and tried to find those of the gestures which (in our opinion) have a vocal counterpart.

Fig. 2. Several examples of the gestures from the Picture dictionary of Gestures [2], which have their vocal counterparts

Tab.1.a to **Tab.1.d** present the list of gestures the scheduled in [2]. Slovak name of the semantic group is shown in first column, and English one in the second column. The characters in the third column represent different types of gestures belonging semantically to the same group. Gestures that have their vocal counterparts are in underlined and bold font.

Tab.1.a

Gestures manifesting physical body		
Veľkosť	*Bigness*	A,B
Chlad	*Coldness*	**A,B,C,D**,E
Hĺbka/ plytkosť	*Depth/ Shallowness*	A
Pitie	*Drink*	**A**,B
Plný nápoja	*Full of drink*	A
Jedenie	*Food*	A,B
Plný jedla	*Full of food*	**A**
Tučnota	*Fatness*	A,B
Horúčka/horúci	*Heat/hot*	**A,B**,C,**D**
Výška	*Height*	A
Dĺžka/Krátkosť/Šírka/ Úzkosť(miera)/Hrúbka/ Tenkosť	*Length/Shortness/Breadth/ Narrowness/Thickness/ Thinness*	A
Spanie	*Sleep*	**A,B**
Štíhlosť	*Slimness*	A,B
Vôňa	*Smell*	**A,B**
Unavenosť	*Tiredness*	A,**B,C**,D
Čo?	*What?*	A

Tab.1.b

Gestures manifesting initiative contact		
Pozornosť	*Attention*	**A**,B,**C**,D,E,**F,G,H,I**,J
Byť ticho	*Be quiet*	**A**,B,C,D,**E**
Prísť	*Come*	**A,B,C,D,E,F,G**
Smery	*Directions*	A,B,C,D,E,F
Pozdravy	*Greetings*	**A,B**,C,D,E,F
Zbohom	*Good-Bye*	**A**,B,C,D,E,F,G,H
Zhon, Rýchlo	*Haste/Quick*	**A,B**
Stop	*Hitchhiking*	**A,B**
Odísť	*Leave*	A,B,C,D
Peniaze	*Money*	A,B,C,D,E,F
Sľub(Prísaha)	*Oath*	A,B,C
Fotografovanie	*Photograph*	**A,B,C**
Obdržať	*Receive*	**A**,B
Opakovanie	*Repetition*	**A,B**
Požiadavky	*Requests*	A,**B,C**,D,**E,F**,G,**H**,I,J
Reštauračné gestá	*Restaurant gestures*	A,B,C,D,E,F,**G**,H,I,J,K,L

Sadnúť	*Sit*	**A**
Spomaliť	*Slow down*	**A**
SMS (Short Message Service)	*SMS*	A
Vstať	*Stand up*	**A**
Zastaviť	Stop	**A,B,C**,D
Telefonovať	*Telephone*	A,B
Čas	*Time*	A,B,C
Čakať	*Wait*	**A,B**
Varovanie	*Warning*	**A,B,C**

Tab.1.c

Gestures manifesting emotional body		
Hnev	*Anger*	**A,B,C,D,E,F,G**,H
Očakávanie	*Anticipation*	**A,B,C**
Ospravedlnenie	*Apology*	**A,B,C,D,E,F,G,H**
Odpor	*Aversion*	**A,B**
Odsúdenie(Zavrhnutie)	*Condemnation*	**A**
Sklamanie	*Disappointment*	**A,B,C**
Rozpaky	*Embarrassment*	**A,B,C,D,E**
Vzrušenie	*Excitement*	**A,B,C,D**
Faux pas	*Faux pas*	**A,B,C,D,E,F**
Strach	*Fear*	**A**
Sklamanie	*Frustration*	**A,B,C**
Vďačnosť	*Gratitude*	**A,B,C**
Šťastie	*Happiness*	**A**
Nenávisť	*Hatred*	**A**
Nedočkavosť	*Impatience*	**A**
Urážky (útoky)	*Insults*	**A,B,C,D,E,F,G,H,I,J**
Bozk	*Kiss*	**A**
Nervozita	*Nervousness*	**A,B,C,D**
Ľútosť	*Regret*	**A,B,C,D**
Pomsta	*Revenge*	**A,B,C,D**
Smútok	*Sadness*	**A**
Sexuálny	*Sexual*	A,B,C
Zahanbenosť	*Shame*	**A**
Prekvapenie	*Surprise*	**A,B,C,D**
Hrozba (vyhrážka)	*Threat*	**A,B,C,D**,E,F,G,**H**

Tab.1.d

Gestures manifesting mental body		
Súhlas(dohoda)	*Agreement*	**A,B**
Súhlas(uznanie)	*Approval*	**A,B,C,D,E,F,G,H**
Komplikácia	*Complication*	**A,B,C,D**
Odmietnutie(popretie)	*Denial*	**A,B,C,D**
Nesúhlas(nezhoda)	*Disagreement*	**A,B,C,D,E**

Neschválenie(nesúhlas)	*Disapproval*	**A**
Úspech	*Good luck*	**A,B,C,D,E,F**
Reči	*Gossip*	**A**
Inteligencia	*Intelligence*	**A**
Nedostatok inteligencie	*Lack of intelligence*	**A,B,C,D,E**
Viacmenej	*More or less*	**A**,B
Žiadna informácia	*No information*	**A**
Dokonalé	*Perfect*	**A,B,C**,D
Modlitba	*Prayer*	**A,B**,C,D
Hrdosť(pýcha)	*Pride*	**A,B,C,D,E,F**,G
Svätosť	*Saintliness*	**A**,B
Seba(Vlastný)	*Self*	**A,B**
Úspech	*Success*	A,**B**,C,**D,E,F,G,H,I,J**
Rozmýšľanie	*Thinking*	**A,B,C,D,E,F**
Víťazstvo	*Victory*	**A,B,C**

It can be seen from the tables, that nearly all gestures manifesting emotional and mental body have their vocal counterparts.

We are aware of the fact, that the vocal forms of the body gestures can confirm the identicalness of the use of the same body gesture for different meanings or on the contrary the same body gesture can have different accompanying vocal gestures that make its meaning different.

From dialogue research most of the vocal gestures are known as discourse markers and backchannels; in speech databases design they are generally classified as speaker noises, grunts, fillers or filled pauses. We also include laughing, crying, and yawning in the set of voice displays that are object of our research.

From the grammatical point of view most of them can be classified as particles or interjections.

Orthographic transcription

A lot of effort has been done to find a set of words that would be able to represent or "encode" orthographically the non-lexical English speech sounds. The most sophisticated sets were designed for speech databases annotation in dialogue and generally speech communication research (e.g. [4], [6]).

The most comprehensive list of orthographic forms of the NVG in Slovak we have obtained from the tools for crossword puzzle solvers and scrabble players and this was enriched by the list of interjections and particles taken from the Slovak National Corpus. The result is a list of several hundreds of orthographic expressions, from which several tenths are candidates to represent speech gestures in our future scheme.

While Picture Dictionary of Gestures offers the picture of the gesture, the name of its class, generic meaning, description (of the body gesture), its specific message in four cultures, and a description of four categories of context (formality, gender, age, social status), the text sources provide us only with orthographic transcription of the sound and in some cases with their linguistic context.

The list will be regularly updated by new gestures found during the database annotation (feedback).

Orthoepy - phonetics

Ordinary phonetic transcriptions of NVGs reflect only the linguistically informative quality. Orthographic transcription is insufficient – it is highly ambiguous and reflects the phonetic content only to very limited extent. Therefore some kind of „allophone transcription“ is needed.

As the standard set of 54 Slovak phonemes and allophones is not sufficient to express the phonetic structure of NVGs, there is a need for new allophones, e.g.:

Glottal stop [ʔ] - which can be used to represent a **hard onset.**

Neutral wovel (schwa) - [ə] - (which does not belong to a Slovak phoneme set)

long neutral wovel (long schwa) - [əː].

Ingressive allophones – e.g. in the word “Hí!” [hiː↓] (IPA character “↓” indicate ingressive airflow).

Ingressive “t↓t↓t↓” (which is in fact not inhaled, because the air is not aspirated by lungs, but sucked in the mouth cavity by the piston function of the tongue. In IPA called dental click marked as [ǀ].

Other ingressive allophones – e.g. in inhaled “Aj jaj ja ja jaj …” [a̯↓j↓ɟ↓a̯↓j↓j↓a̯↓j↓a̯↓ɟ↓a̯↓j↓] or “Ej jej jéj ...” [e̯↓j↓ɟ↓e̯↓j↓ɟ↓e̯ː↓j↓].

New nasal allophones:

nasal murmured [ɦ̩] e.g. in the grunt “mhm” [ʔm̥ɦ̩ᵖm̥] pronounced with closed mouth,

nasal schwa = long m - [m̥ː],

long n - [n̥ː]

New allophones (consonants) to create syllable nucleii (e.g. to use in the words imitating sounds: [ps̩ːt], [ne-ps̩ː-kaj] (= Don’t make “psss” sound!)):

[s̩ː], [sʲ̩ː], [z̩ː], [zʲ̩ː], [x̩ː], [m̥ː], [n̥ː], [f̩ː] etc.

New quantity class should be added – extra long allophones: e.g. “Áááááh” should be transcribed like [aːːh].

Terminology ☺

The allophones we are speaking about in fact do not follow exactly all aspects of the classical definition of allophone.

How about calling them:	NVG allophones	→ nallophones	
The same holds with:	NVG syllables	→ nyllables	
	NVG words	→ nords	
Or even:	NVG phrase	→ nrase	...and
	NVG sentence	→ nentence	

Prosody

There is a deeper relatedness between the verbal and the non-verbal than is generally understood. It seems probable that pre-verbal archetypes of vocal gestures exist, and that the newer versions of gestures keep similar prosody to their ancient predecessors. Note the similarity among the pitch contours in the interjection "íha" and its semantic equivalent "mm-hm" (Fig. 3a). The same phenomenon can be observed on the "čo?" – "hm?" (= *what?)* pair and many others. Also some syllabic/rhythmic patterns seem to be adopted to lexical words from their non-lexical counterparts. Slovak word "nie" (= *no*) is normally pronounced as one syllable with a diphthong [i̯e] in its nucleus. But in expressive speech with strong emphasis on this word it would be often pronounced as "ni-je" [nʲi i̯je], broken into two syllables in the same way as its non-lexical counterpart "ə-ə" [ʔəʔə] (= *no*) (Fig. 3b).

We are aware of complexity of these dependences and relations which can illustrated on many examples that do not follow the mentioned principals (Fig 3c, showing "áno" – "mhm" [ʔm̩ɦiᵖm̩] pair expressing positive answer (= *yes*), but having totally different pitch contours).

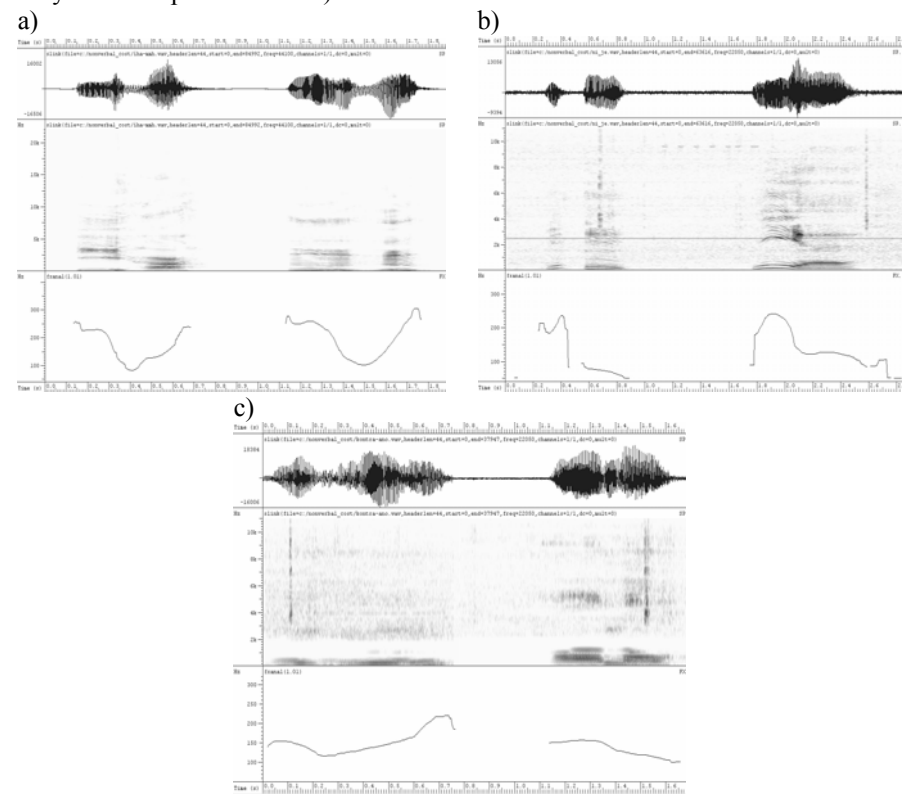

Fig. 3. Oscillograph, spectrograph and pith contour of the examples described higher a) "íha" - "mm-hm" [m̩:ɦiᵖm̩], b) "nie" - "ə-ə" [ʔəʔə], "áno" - "mhm" [ʔm̩ɦiᵖm̩].

Intonation model

The frequency range in expressive speech can be enlarged and 2-3 levels of tones high, (middle), low (H, (M,) L), generally used for standard speech may need to be expanded to 4-5 levels (HH, H, (M) L, LL). For instance in the word „Really?“ the pitch reaches L,H tones in neutral speech, LL,L when surprised&having doubt, H,HH when excited.

Moreover the pitch contour is often changing fluently (gliding) or in small steps. Therefore we think that basic set of tone symbols that would give a usable simplified representation of pitch contour of the NVGs can be: T – very high tone (top), H – high tone, L – low tone, B – very low tone (bottom), D – small step downwards (e.g. “No, no, no, no, no, no, no!” – T D D D D D D L), U – small step upwards, and their “glided” combinations (excluding D and U) to express continual rise or fall: H+T, L+T, L+B, etc.

Voice quality – Loudness – Timing

Voice quality is a very important feature in expressive speech and it carries a high amount of information. Some authors consider voice quality to be “the fourth prosodic parameter” [7].

Although voice quality can be reflected to a certain degreee in the definition of the “noneme” set, it is better to use Laver’s voice quality classification scheme to annotate strongly changed (non-standard) voice quality.

More complex NVGs seem to be organized in time in syllable-like segments. Sometimes they are shorter (in a *staccato* manner, filling only part of the “beat”), other time they are longer and not separated by any pauses (which resembles notes played *tenuto*).

Therefore we would like to allow using terms/signs from Italian musical terminology (*pp, ff, stacc.* etc.) to mark some phenomena of loudness and rhythm. These signs would serve only as a comment.

Semantics and Pragmatics

How is it possible, that a single vocal gesture is often able to bear the same information as the whole sentence? One of the possibilities is that the gesture occurs often in a fixed combination with certain type of semantic units (SU) with certain meaning. Then when the gesture occurs alone, it borrows the same meaning as the NSG-SU combination.

To be able to check hypotheses like this, we will need a kind of semantic annotation system.

Semantics represents the first level, significant meaning, i.e. what the gesture “says” in general. We are inspired by the semantic description and classification of the gestures in body gesture dictionaries (e.g. [2]), by classification schemes of

particles and interjections, and by the different metadata annotation specifications (e.g. [8])

Pragmatics represents the situational meaning - what does the speaker really mean/intend by his gesture shown to the particular interlocutor in the actual situation (second level meaning). Differences in pragmatic meaning are often achieved by adding ironical, satirical undertone etc.

The communication follows certain rules based on conventions and it will be successful only if the receiver understands both meanings (the standard one, and the situational one). Therefore the gestures are culturally dependent and they can be exported/imported together with foreign culture (Yessss! Wow!).

Conclusion

A preliminary stage of the research of the nonverbal speech gestures in Slovak is presented. The authors discuss their ideas of the development of a coding scheme which would make the annotation of nonverbal vocal gestures in Slovak possible.
Topics discussed in the paper include orthographic annotation, orthoepic representation using extended allophone set, intonation coding scheme and the need for new terminology. Voice quality, loudness, timing, semantic and pragmatic content annotation, were also discussed.

Acknowledgments. This work was supported by the of the Ministry of Education of the Slovak Republic, Scientific Grant Agency project number 2/0138/08 and Applied Research project number AV 4/0006/07.

References

1. http://korpus.juls.savba.sk/
2. Ružičková, E.: Picture dictionary of gestures (American, Slovak, Japanese, and Chinese), Comenius University Publishing House, Bratislava (2001), ISBN 80-223-1675-X
3. Eckert, H., Laver, J.: Menschen und Ihre Stimmen., Beltz Psychologie Verlags Union, Weinheim, (1994)
4. Ward, N.: Non-Lexical Conversational Sounds in American English., Pragmatics and Cognition, 14:1, (2006) 113-184
5. http://en.wikipedia.org/wiki/Gesture
6. http://www.univie.ac.at/voice/documents/VOICE_spelling_conventions_v2-1.pdf
7. Campbell, N, Mokhtari, P., Voice Quality; the 4th prosodic parameter, in Proc 15th ICPhS, Barcelona, Spain, 2003.
8. http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf