# INFLUENCE OF VERBAL CONTENT ON ACOUSTICS OF SPEECH EMOTIONS

Hille Pajupuu, Jaan Pajupuu, Kairi Tamuri, Rene Altrov

Institute of the Estonian Language, Tallinn, Estonia
hille.pajupuu@eki.ee, jaan@pajupuu.eu, kairi.tamuri@eki.ee, rene.altrov@eki.ee

**ABSTRACT**

This paper deals with the issue of the influence of verbal content on listeners who have to identify or evaluate speech emotions, and whether or not the emotional aspect of verbal content should be eliminated. We compare the acoustic parameters of sentences expressing joy, anger, sadness and neutrality of two groups: (1) where the verbal content aids the listener in identifying emotions; and (2), where the verbal content does not aid the listener in identifying emotions. The results reveal few significant differences in the acoustic parameters of emotions in the two groups of sentences, and indicate that the elimination of emotional verbal content in speech presented for emotion identification or evaluation is, in most cases, not necessary.

**Keywords**: acoustic features, emotional speech, emotion identification, verbal content.

## 1. INTRODUCTION

Research of emotional speech and its application in speech technology presumes speech material where the emotional colouring has previously been identified by listeners (evaluators). The tests are arranged so that the listeners have to judge on the emotion from the voice only, minimizing the influence of the verbal content[1] of the heard text.

As early as 1964 Kramer stated that "judging emotion from the nonverbal properties of speech requires elimination of verbal cues" [10]. He proposed three methods how to eliminate the verbal influence: (a) use a constant, ambiguous set of words for various emotional expressions; (b) filter out the frequencies which permit word recognition; (c) use listeners who do not know the language. Additional methods used to eliminate the possible effect of verbal influence on emotion judgement include (d) random spliced speech [17], (e) reversed emotional speech / backward speech [19] and (f) speech-like but meaningless sentences uttered with different emotions [18].

Over the recent decade, research on speech emotions has received a considerable impetus owing to the rise of emotional speech corpora and to the application prospects offered by speech technology. Several studies still use testing material where the influence of the verbal content on emotion identification is avoided, see, for example, [6, 7, 12, 16, 20]. Of the above methods, manipulation of the speech signal (random spliced speech, reversed speech) is less used, as it may destroy part of the information necessary for emotion perception. As for filtering, there is still no consensus as to what is the optimal filter for removing the verbal content without losing a considerable part of the acoustic emotional content [9]. Neither is a lot of hope put on evaluators from other cultural and linguistic backgrounds: newer results have revealed not only universal traits in the vocal expression of emotions, but also some cultural specifics, which may fail the cross-cultural decoding of emotions [1, 8, 13]. The available material such that is indeed free from verbal cues has been obtained for the corpuses and databases of emotional speech mainly by having actors read neutral or meaningless sentences with various acted emotions, see the surveys [3, 4, 14].

The current trend of studying and processing natural emotional speech (emotion recognition, synthesis of emotional speech) has raised a need for a different kind of emotional speech material, preferably drawn from speech with spontaneous or elicited emotions [5, 11, 15, 23, 24]. While collecting material for emotional speech corpora using natural speech, one can hardly expect to find two similar sentences which have a neutral content and yet have been uttered with different emotions.

In order to enable emotion identification in natural speech samples without letting the verbal content disturb the listeners' judgement on the speech emotion one could, of course, return to word filtering. Low-pass filtering ($\leq$ 1000 Hz), for example, would be conceivable as in this case the speech signal will at least retain some information that is necessary for emotion perception [21, 22]. However, one may ask whether elimination of the effect of the verbal content is necessary at all after this change of goals.

The aim of our study was to find out whether there is any acoustic difference between two groups of sentences: (1) those where the verbal content affects the listeners in their attempts of emotion recognition and (2) those where it does not.

The research question was posed as follows: Is it necessary to eliminate the influence of the verbal content from the sentences presented to listeners for the identification of vocal expression of emotions?

## 2. MATERIAL AND METHOD

The material comes from the Estonian Emotional Speech Corpus[2]. As far as we know, this is the only freely available corpus of emotional speech where the material has been divided into two groups: Sentences the verbal content of which has been found to affect the listeners in emotion identification, and sentences where no such influence has been detected. The corpus contains 1,234 sentences read by a female voice. The sentences (all different) have been extracted from longer recorded text passages. The reader has not been instructed to use any particular emotion, assuming that any text evokes a certain mood sounding in the reading voice. Thus, the emotions are not acted, but elicited by the text, while the emotions are not expressed in full, but rather moderately. To provide the sentences with an emotion label the sentences were separated from the context and presented to a group of evaluators (14 web-based listening tests, for each test about 30 adult listeners whose native language was Estonian) who were asked to decide whether the sentence sounded joyous, angry, sad, or neutral. The listeners were briefed that each of the three emotions listed (joy, anger, sadness) also comprised several other closely related emotions: joy included gratitude, happiness, pleasure and exhilaration; anger included resentment, irony, reluctance, contempt, malice and rage; sadness covered loneliness, disconsolation, concern and hopelessness, while neutral speech was to be understood as normal speech, without special emotions. The sentences could be listened to as many times as needed. For 73.5% of the corpus sentences the same emotion or neutrality was suggested by over 50% of the listeners (i.e. more than 2 times better than chance), see [2].

In addition, the same sentences were presented to another group of evaluators (in 14 reading tests), who had not participated in the listening tests. They were asked to decide on the emotion from the written text, without hearing the sound. Based on a comparison of the listening and reading results the corpus sentences were divided into two groups: (1) Sentences where the verbal content conveys a similar emotion as the voice (a similar emotion has been identified both by readers and listeners). For example, the sentence ■ *Täiesti mõistetamatu!* [*Completely incomprehensible!*] has been classified under anger by 83.0% of the readers and by 100% of the listeners. (2) Sentences where the tone of the voice changes the emotion detected from the purely verbal content (the reading and listening groups have identified different emotions). For example, the sentence ■ *Ükskõik, mida ma teen, ikka pole ta rahul!* [*Whatever I do, he is never satisfied!*] has been classified under anger by 64.3% of the readers, whereas 80.0% of the listeners have decided on sadness. In addition, this group contains sentences where the verbal content fails to reveal an unambiguous emotion and thus it is the voice that makes all the difference. For example, the sentence ■ *Ehkki Ott minu olemasolust midagi ei teadnud.* [*Although Ott knew nothing of my existence.*] has not received an unambiguous emotion identification from the readers, but 87.5% of the listeners have found it to be joy; for details, see [2]. The sentences in Group 1 are such where the verbal content of the message may help the listeners to decide on the emotion. Group 2, however, contains sentences where the listeners decide on the emotion by the voice, not the verbal content.

For the present study, the corpus was searched for sentences where anger, joy, sadness or neutrality had been identified by at least 51% of listeners (more than 2 times better than chance probability), see Table 1.

**Table 1**: Material: number of sentences.

| Groups | Neutral | Anger | Joy | Sadness |
|---|---|---|---|---|
| Verbal influence | 95 | 158 | 171 | 94 |
| No verbal influence | 103 | 79 | 60 | 87 |
| Mean percentage of identification by listening and std | 68.3<br>11.9 | 73.3<br>14.6 | 75.4<br>14.5 | 72.1<br>14.7 |

We focused on three features which are most consistently used in the research of speech emotions: loudness (intensity), pitch (F0), and duration (cf. [16]).

Intensity Parameters (dB):
- Intensity mean
- Intensity range (difference between $Intensity_{max}$ and $Intensity_{min}$)
- Intensity start (intensity at sentence onset)
- Intensity end (intensity at sentence end)

F0 Parameters (Hz):
- F0 mean
- F0 range (difference between $F0_{max}$ and $F0_{min}$)
- F0 start (pitch at sentence onset)
- F0 end (pitch at sentence end)

Duration Parameters (ms):

Mean vowel durations.

The difference between the groups was determined by the Wilcoxon rank sum test.

First, for all emotions, each acoustic parameter was compared with its counterpart in neutral sentences of the same group ('verbal influence' or 'no verbal influence').

The same test was applied for an inter-group (verbal influence – no verbal influence) comparison of the parameters of similar emotions (neutral – neutral, anger – anger, joy – joy, sadness – sadness).

Also, the Wilcoxon rank sum test was used to compare the groups of sentences classified under neutral, anger, joy, and sadness, in order to find out whether the given parameters still distinguish between the emotions, if the material with verbal influence is pooled with that with no verbal influence.

## 3. RESULTS

Table 2 presents the median values of the parameters of the neutral sentences of two groups – (1) the verbal content influences emotion identification and (2) the verbal content does not influence emotion identification – and increase or decrease of the value of the emotional sentences as compared to the neutral ones. Also, for each parameter of these two groups the p-value for Wilcoxon rank sum test has been calculated for all emotions. The difference between the two groups appeared to be significant only for some of the parameters and some of the emotions. The intensity mean significantly distinguished the two groups in the case of neutrality ($p$=.001), anger ($p$=.001) and joy ($p$=.024). The intensity range was significantly different in sadness ($p$=.008). The F0 mean also differentiated between the two groups in sadness ($p$=.001) and neutrality ($p$=.021), while duration did it in the case of joy ($p$=.001).

Table 3 presents the $p$-values of Wilcoxon rank sum test for all emotions analysed. The results show that the $p$-values remain significant even though we do not distinguish between the groups where those emotions have been identified with the help of the verbal content of the sentence or without it. Intensity mean, F0 mean, F0 range and duration turned out to be the best distinctive features for the emotions investigated.

Table 2: Comparison of the acoustic parameters of emotions as manifested in two sentence groups (verbal influence present vs. absent in emotion identification)

| | Verbal influence | Neutral | Anger | Joy | Sadness |
|---|---|---|---|---|---|
| **Intensity mean** | Yes | 70.9 | > | <*** | <*** |
| | No | 71.5 | <** | < | <** |
| Wilcoxon rank sum test | *p* | **.001** | **.001** | **.024** | .405 |
| **Intensity range** | Yes | 14.7 | > | < | < |
| | No | 13.7 | > | < | > |
| Wilcoxon rank sum test | *p* | .465 | .791 | .539 | **.008** |
| **Intensity start** | Yes | 74.2 | < | < | < |
| | No | 74.8 | < | < | <** |
| Wilcoxon rank sum test | *p* | .234 | .725 | .598 | .912 |
| **Intensity end** | Yes | 65.5 | > | < | < |
| | No | 65.1 | < | < | <* |
| Wilcoxon rank sum test | *p* | .815 | .240 | .890 | .314 |
| **F0 mean** | Yes | 183.5 | <** | >*** | <* |
| | No | 185.4 | <*** | > | = |
| Wilcoxon rank sum test | *p* | **.021** | .966 | .926 | **.001** |
| **F0 range** | Yes | 91.7 | >*** | >* | <* |
| | No | 94.7 | >* | > | < |
| Wilcoxon rank sum test | *p* | .405 | .830 | .595 | .221 |
| **F0 start** | Yes | 227.0 | <* | < | <* |
| | No | 232.0 | < | < | <* |
| Wilcoxon rank sum test | *p* | .344 | .123 | .423 | .827 |
| **F0 end** | Yes | 163.6 | < | > | > |
| | No | 165.4 | < | < | < |
| Wilcoxon rank sum test | *p* | .548 | .989 | .260 | .548 |
| **Vowel duration mean** | Yes | 66.0 | < | >*** | > |
| | No | 65.0 | <* | < | < |
| Wilcoxon rank sum test | *p* | .210 | .700 | **.001** | .158 |

Note. *Presented are the median values of the neutral parameters. The left bracket < marks a decrease from the neutral value, the right bracket > marks an increase from the neutral value, * at the angle bracket marks a significant difference of the emotion from the neutral value: * p< .05, ** p< .01, *** p< .001. The last line for each parameter group contains the p-value of Wilcoxon rank sum test between groups with the same emotion but different verbal influence. Significant differences are emphasized in bold. The intensity was measured in dB, the pitch (F0) in Hz, and duration in ms.*

**Table 3**: The *p*-value of Wilcoxon rank sum test between all pairs of the four sentence groups (neutral, anger, joy, sadness) by parameters.

| | Anger | Joy | Neutral |
|---|---|---|---|
| **Intensity mean** | | | |
| Joy | .001*** | | |
| Neutral | .051 | .001*** | |
| Sadness | .001*** | .001*** | .001*** |
| **Intensity range** | | | |
| Joy | .018* | | |
| Neutral | .184 | 1.00 | |
| Sadness | .069 | 1.00 | 1.00 |
| **Intensity start** | | | |
| Joy | .758 | | |
| Neutral | .758 | .282 | |
| Sadness | .001*** | .006** | .001*** |
| **Intensity end** | | | |
| Joy | .096 | | |
| Neutral | .411 | .391 | |
| Sadness | .001*** | .096 | .005** |
| **F0 mean** | | | |
| Joy | .001*** | | |
| Neutral | .001*** | .001*** | |
| Sadness | .001*** | .001*** | .029* |
| **F0 range** | | | |
| Joy | .027 | | |
| Neutral | .001*** | .010** | |
| Sadness | .001*** | .001*** | .002** |
| **F0 start** | | | |
| Joy | .167 | | |
| Neutral | .006** | .358 | |
| Sadness | .781 | .094 | .001*** |
| **F0 end** | | | |
| Joy | .001*** | | |
| Neutral | .021* | .127 | |
| Sadness | .040* | .127 | .891 |
| **Duration** | | | |
| Joy | .001*** | | |
| Neutral | .001*** | .003** | |
| Sadness | .001*** | .021** | .558 |

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

## 4. DISCUSSION

Often, studies of emotional speech avoid material where emotion identification might be influenced by the verbal content of the test sentences [6, 7, 12, 20]. The present study was meant to find out whether there is a significant difference between the acoustic parameters of moderately expressed emotions in sentences of two groups – one where emotion identification has been influenced by the verbal content, and the other where it has not. According to our results, the inter-group difference was significant only for some parameters and some emotions (see Table 2). This suggests that efforts to eliminate verbal influence from sentences presented to evaluators for emotion identification might perhaps as well be spared.

Next we used the pooled material of the two groups to investigate whether significant acoustic differences can be established between the joyous, angry, sad, and neutral sentences. The differences appeared to be significant indeed (Table 3), which suggests that emotion studies need not be confined to the material from which the verbal content influence has been eliminated.

Elimination of the verbal influence was certainly necessary at the time when it was not yet clear if emotions were identifiable from the voice at all, and also, when research material was scarce and easy comparability was an issue. At that time actors were used to produce the emotional speech, and the study of the emotions in natural speech was a subject for the future. As the material thus collected is available in the emotion corpora it is still used, even for training emotion classifiers, although their subsequent target will be natural speech. In natural speech, emotions may be conveyed by both lexical and acoustic means, or sometimes exclusively by the latter.

As success in emotion recognition depends on the similarity of the training material with the target speech [8, 23], corpora of emotional speech should obviously contain material occurring in natural speech. Thus, evaluation of the corpus material should proceed without worrying about the influence that the verbal content might exert on the evaluator, as in real life we decode speech emotions despite its "disturbing" effect anyway.

## 5. CONCLUSION

Our investigation of the influence of the verbal content of sentences on the acoustics of speech emotion has demonstrated that for speech emotion research, there is no significant difference between the speech material where the verbal content can influence emotion recognition and the material where it cannot. Therefore, evaluation of corpus material containing natural speech emotions can be done without tackling the problem how to eliminate verbal influence.

## 6. ACKNOWLEDGEMENTS

The study was supported by the institutional research funding IUT35-1 of the Estonian Ministry of Education and Research and the Alfred Kordelin Foundation.

## 7. REFERENCES


[1] Altrov, R. 2013. Aspects of cultural communication in recognizing emotions. *Trames: Journal of the Humanities and Social Sciences*, 17 (67/62), 159-174.

[2] Altrov, R., Pajupuu, H. 2012. Estonian Emotional Speech Corpus: Theoretical base and implementation. In: Devillers, L., Schuller, B., Batliner, A., Rosso, P., Douglas-Cowie, E., Cowie, R., Pelachaud, C. (eds). *4th Int. Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3)*, Istanbul, 50-53.

[3] Bänziger, T., Mortillaro, M., Scherer, K. R. 2011. Introducing the Geneva Multimodal Expression Corpus for experimental research on emotion perception. *Emotion*, 12 (5), 1161-1179.

[4] El Ayadi, M., Kamel, M. S., Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44 (3), 572-587.

[5] Fernandez, R., Picard, R. 2011. Recognizing affect from speech prosody using hierarchical graphical models. *Speech Communication*, 53 (9-10), 1088-1103.

[6] Grichkovtsova, I., Morel, M., Lacheret, A. 2012. The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54 (3), 414-429.

[7] Jaywant, A., Pell, M. D. 2012. Categorical processing of negative emotions from speech prosody. *Speech Communication*, 54 (1), 1-10.

[8] Kamaruddin, N., Wahab, A., Quek, C. 2012. Cultural dependency analysis for understanding speech emotikon. *Expert Systems with Applicat*, 39 (5), 5115-5133.

[9] Knoll, M., Uther, M., Costall, A. 2009. Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners. *Speech Communication*, 51 (3), 210-216.

[10] Kramer, E. 1964. Elimination of verbal cues in judgments of emotion from voice. *J. of Abnormal and Social Psychology*, 68 (4), 390-396.

[11] Mori, H., Satake, T., Nakamura, M., Kasuya, H. 2011. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53 (1), 36-50.

[12] Patel, S., Shrivastav, R. 2011. A preliminary model of emotional prosody using multidimensional scaling. In: *Interspeech*, Florence, 2957-2960.

[13] Pell, M. D., Monetta, L., Paulmann, S., Kotz, S. A. 2009. Recognizing emotions in a foreign language. *J. of Nonverbal Behavior*, 33 (2), 107-120.

[14] Pittermann, J., Pittermann, A., Minker, W. 2010. Handling emotions in Human-Computer Dialogues. Dordrecht–Heidelberg–London–New York: Springer.

[15] Polzehl, T., Schmitt, A., Metze, F., Wagner, M. 2011. Anger recognition in speech using acoustic and linguistic cues. *Speech Communication*, 53 (9-10), 1198-1209.

[16] Roche, J. M., Peters, B.,Dale, R. 2015. „Your Tone Says It All“: The processing and interpretation of affective language. *Speech Communication*, 66, 47-64.

[17] Scherer, K. R. 1971. Randomized splicing: A note on a simple technique for masking speech content. *J. of Experimental Reseach in Personality*, 5 (2), 155-159.

[18] Scherer, K. R., Banse, R., Wallbott, H. G., Goldbeck, T. 1991. Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15 (2), 123-148.

[19] Scherer, K. R., Ladd, R. D., Silverman, K. E. A. 1984. Vocal cues to speaker affect: Testing two models. *J. Acoust. Soc. Am*., 76 (5), 1346-1356.

[20] Scherer, K. R., Scherer, U. 2011. Assessing the ability to recognize facial and vocal expressions of emotion: construction and validation of the emotion recognition index. *J. of Nonverbal Behavior*, 35 (4), 305-326.

[21] Snel, J., Cullen, C. 2011. Obtaining speech assets for judgement analysis on low-pass filtered emotional speech. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition – FGR*, 835-840.

[22] Snel, J., Cullen, C. 2013. Judging emotion from low-pass filtered naturalistic emotional speech. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)*. Geneva, Switzerland, 336-342.

[23] Vogt, T., Andr, E., Wagner, J. 2008. Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. In: Peter, C., Beale, R. (eds), *Affect and Emotion in Human-Computer, Interaction, LNCS 4868*. Heidelberg, Germany: Springer, 75-91.

[24] Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S. 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. on pattern analysis and machine intelligence*, 31 (1), 39-58.


[1] The terminology varies: semantic content, linguistic content, lexical content, textual or text content are also used.

[2] http://peeter.eki.ee:5000/