Emotional Prosody - Does Culture Make A Difference?

F. Burkhardt$^{(1)}$, *N. Audibert*$^{(2)}$, *L. Malatesta*$^{(3)}$, *O. Türk*$^{(4,5)}$, *L. Arslan*$^{(4,5)}$ *& V. Auberge*$^{(2)}$

$^{(1)}$T-Systems Enterprise Services GmbH, Berlin, Germany
$^{(2)}$ICP - University of Stendhal, Grenoble, France
$^{(3)}$IVML - Technical University of Athens, Greece
$^{(4)}$R&D Dept., Sestek Inc., Istanbul, Turkey
$^{(5)}$Electrical and Electronics Eng. Dept., Bogazici University, Istanbul, Turkey

Abstract

We report on a multilingual comparison study on the effects of prosodic changes on emotional speech. The study was conducted in France, Germany, Greece and Turkey. Semantically identical sentences expressing emotional relevant content were translated into the target languages and were manipulated systematically with respect to pitch range, duration model, and jitter simulation. Perception experiments in the participating countries showed relevant effects irrespective of language. Nonetheless, some effects of language are also reported.

1. Introduction

With the proceeding pervasiveness of speech-based human-machine interfaces emotional speech gets more and more into focus, with the development of more natural dialog systems the simulation of emotional speech is desirable. But if we develop an emotional speech synthesizer based on the statistical analysis of natural databases, how can we be sure that the emotional expression is plausible across cultures and languages?

Based on an emotional speech synthesizer available for over 34 languages, MBROLA [8] enhanced by Emofilt [3], we conducted a cross cultural listening experiment to evaluate the effect of prosodic manipulations on emotional impression across languages. The Emofilt program as well as audio samples can be found on the web at <http://emofilt.sourceforge.net>

This paper is structured as follows. The next section discusses preceding work found in the literature. Although already a number of studies deal with the inter cultural aspects of emotional speech, few regarded synthesized speech. Section 3 deals with the question how to denominate emotion related states. In order to lessen the problem of finding "emotion-words" that mean the same across all target languages, we used key phrases that allude to emotional situations. In section 4 we describe the prosodic manipulations we performed on the "neutrally" spoken phrases. They are based on systematic variation. The 5. section describes the procedure of the listening experiment. Finally, in section 6 we report on the results and conclude in section 7 with a final discussion.

2. Related Work

First attempts to simulate emotional speech by means of speech synthesis started soon after the first mature speech synthesizers were developed, e.g. [6] and is gaining rising attention with the more widespread use of speech synthesis in voice-portals, multimodal user interfaces and talking heads. For an overview on the history of emotional speech synthesis the reader may be referred to [12]. One of the most challenging tasks today for speech synthesis is to find solutions for the trade-off between the naturalness of data-concatenation engines and the flexibility of signal generation synthesis. This problem is evident in the case of emotional speech synthesis, where voice quality features or articulatory precision become important.

In the case of diphone-synthesis there exist two approaches to add voice quality control:

- Multiplying the diphone database by variants with different vocal efforts, e.g. [13].
- Modification of the voice quality in real time, e.g. by modifying a LPC residual [5] or sinusoidal modeling [7]

We didn't use such techniques in this investigation as we didn't want to change the original MBROLA engine but applied a very crude jitter-simulation. Therefore, our technique did not require off-line processing of databases as well as the application of intense signal processing algorithms during synthesis.

The fact that intercultural aspects of emotional perception make a difference could be shown e.g. in [9]. One of the outcomes of this study, which used portrayed emotional nonsense speech by actors, was that indeed the only country not belonging to the Indo-European language family achieved significantly worse recognition rates. In another study by Abelin et al [1], the authors worked in the opposite direction and showed that the interpretation of emotions by listeners with different mother-tongues depend on the intended emotions. Specifically anger, fear, sadness, and surprise were interpreted as intended in a greater degree as compared to shyness, dominance, happiness, and disgust by listeners with different native languages. A recent study, [2], which investigated speech samples dubbed from TV-series, also confirmed the cultural differences in coding as well as perception of emotional speech. To our knowledge there was no investigation up to now to tackle the issue with the issue of synthetic speech in mind.

3. Coding "Emotion-Related States"

In the literature the overwhelming number of synthesis experiments regarded a very limited set of so-called "basic" or "fundamental" emotions, well known as "the big four / six". The problem with respect to real world applications is that these pure emotions almost never occur in natural data. We therefore wanted to avoid the explicit naming of emotion-terms and coded the emotion-related states in phrases that might have been uttered in emotionally strongly affected situations. This technique was inspired by Marc Schröder's work [11]. Another advantage of this approach consists of the little cognitive load that the listeners of the perception experiment had to manage; they only had to answer one single question: ”*was the sentence uttered in an appropriate way?*”. Another criterion for the choice of wording was that the sentences were conceivably utterable by a machine in order to distract from the artifacts of the speech synthesizer. The set of sentences chosen to be appropriate for the six target emotion-related states plus neutral are listed in table 1. The emotional states are also annotated with their values of the widely used emotional dimensions activation, valence and dominance. Actually, the original motivation for the choice of emotional states where to target the extreme points in a cube spanned by these dimensions. We admit that these categorisation is disputable and somewhat arbitrary. The usage of these sentences also demonstrates a possible application of simulated emotional speech: to enhance a message's effect by strengthening the semantics on the extra-lingual level.

Table 1: *The emotion related states and their key phrases.*

Target emotion-related state	Dimensional classification in activation-valence-dominance space	English key-phrase
Neutral	-	*You've got seven new messages in your mailbox since yesterday.*
Joyful	aroused, pleasant, dominant	*Congratulations, you've just won the lottery!*
Friendly	calm, pleasant, dominant	*A very warm welcome to our voice portal service.*
Threatening	either, unpleasant, dominant	*You didn't react to our dunning letter, further steps will be taken.*
Bored	calm, unpleasant, neither	*This is the five thousand three hundred second status report. All systems are up and running.*
Frightened	aroused, unpleasant, subdominant	*The brake-system reports a severe malfunction.*
Sad	calm, unpleasant, subdominant	*Agent b-thirty five's life-functions have ceased 5 minutes ago.*

4. Prosodic Manipulation

As the focus of this investigation was on the lingual and cultural universality of prosodic manipulation with respect on the emotional impact, we couldn't use ”emotion-rules” like mentioned in [6] or [4], because all of these rules are derived from data that is (possibly) only valid for a specific language. Instead we did a systematic variation on the three parameters pitch range, duration, and jitter. The following lists the parameters and remarks on the expectations with respect to their effect on the emotional dimensions.

- Pitch range variation among narrow, original, and broad. The hypothesis was, that phrases that with high activity connotation like joy or anger would sound more natural with a broader range, while those with low arousal like boredom or sadness with a smaller range. Also pleasant emotions might be better represented by a broader range and vice versa.
- Duration variation among slower, original, and ”stressed” (everything faster, accents slower). The hypothesis would say that a low arousal goes along with slow speech rate and urgent messages like the frightened or the joyful one are better represented by a stressed manner of speaking.
- Jitter simulation varying between original and Jitter (raising and lowering each pitch-value by 10 percent in a 10 msec distance). The jitter should represent emotions with negative valence like sadness or fear, as it gives the voice a somehow ”crying” effect.

Considering the seven target phrases we prepared 3 * 3 * 2 * 7 = 126 sentences in total. This amount can be assessed by listeners in one session (about 15 minutes) without presenting unduly labor.

For the manipulations, the Emofilt-software [3] was used, which is a frontend for the MBROLA speech synthesizer [8]. MBROLA is a diphone synthesizer with databases for over 34 languages in many voices. It accepts as input format a list of phoneme designators aligned with a prosodic description that consists of a duration value and a set of target pitch values.

Emofilt is simply a program to apply rules like “*raise pitch by 50 %*” with this input format. Since the rules always take an ”emotionally neutral version” as the reference, all the partners translated and recorded the target phrases in their mothertongues in a neutral, ”matter-of-fact” style and extracted the phonetic and prosodic description manually to be used as a reference for the Emofilt-program. Using Emofilt and MBROLA, the test-audio files were then generated automatically and identically for all the languages investigated. This means the same rules were applied to all languages, but of course they differed with respect to their effect, as the neutral references were not the same. We used the male databases de6, fr6, gr2 and tr1.

5. Perception Experiment

The stimuli were judged by ten female and ten male listeners in each country, every listener being a native speaker of the tested language, without any known hearing disorder. The age distribution was between 16 and 58 with mean value 30 and standard deviation 8.2. All listening experiments were done using the same automated test-program. Each participant gets her/his own random order and can listen to the stimulus only once. The question to be answered was: ”*On a scale from 1 to 7, how appropriate is the manner of speaking for the meaning of the sentence? (1 = does not fit at all - 7 = fits fairly well)*”. The tests were done in quiet surrounding on a laptop with headphones.

6. Results

We computed a four factorial analysis of variances with complete repetition of measurement on the data using the SPSS statistical software. The four inner subject factors were sentence, pitch, duration, and jitter. The three between-subject factors were language, gender, and age. Whereas there were no significant effects for the between-subject factors, there were some for language in conjunction with the inner subject variables.

6.1. Main Effects

The sentence alone showed a significant effect (all results reported showed significance better than .05 if not otherwise stated). In general, it seems that the listeners found the negative emotions better displayed by the speech synthesizer than the positive ones. This might be due to the fact that most listeners were not used to listening to diphone synthesis and found that a machine-like manner of speaking suits better for an unpleasant message.

The other inner-subject factors also showed main effects. The results show that phrases with the original prosody were generally judged as more appropriate then those that were manipulated, irrespective of the sentence. This showed especially for broad range, slow speaking rate and jitter-simulation. It seems that the manipulations always sound somehow unnatural.

6.2. Effects Depending on Pitch-range and Duration

Although pitch-range and duration modification showed a significant effect for the different phrases individually, the combined effect was also significant and we will confine on commenting on these results.

Neutral phrase
The neutral phrase was clearly best represented by the original prosody. From this result it can mainly be derived that the modifications were clear enough to be detectable for the listeners as a style of speaking not appropriate for non-emotional communication.

Joyful phrase
The phrase that indicated a joyful situation was, according to the hypotheses derived from literature, best represented by a broad range and stressed manner of speaking. The opposite manipulations, slow speech and small range, were clearly not appropriate for joyful meaning. We see that the manipulations worked indeed as anticipated, irrespective of language and culture.

Friendly phrase
The phrase carrying a friendly message was equally well represented by a normal pitch range with neutral or stressed durations. It can be seen similar to the neutral phrase, but the broad pitch-range and stressed duration-model is almost equally adequate. This fits nicely if you regard the friendly message as a somewhat damped version of the joyful one.

Threatening phrase
The threatening phrase was well represented by a small pitch-range and stressed durations. Clearly a broad pitch-range was not adequate for the threatening message which gets along with the predictions found for ”cold anger” in the literature ([10], p. 158).

Boring phrase
The results for the phrase meant to carry a boring message do not show a clear structure; it seems that the listeners didn’t interpret the message in the expected way.

Frightened phrase
The results for the frightened phrase look somewhat similar to the threatening one in the sense that small range and stressed durations are most appropriate with the difference that a normal way of speaking is less adequate for frightened speech then for threatening.

Sad phrase
The phrase carrying a sad message was, in conformance with the hypothesis, best represented with a small pitch-range. Although the slow duration-model was not the preferred one, it is clearly much higher rated then for the other sentences.

6.3. Effects of Jitter-Simulation

Because the jitter-simulation is very crude and doesn’t sound natural, it was never really rated better then the unmodified versions. Nonetheless, for sadness and fear this effect is, as predicted by the literature, considerably smaller which shows that the intended effect was perceived by the listeners.

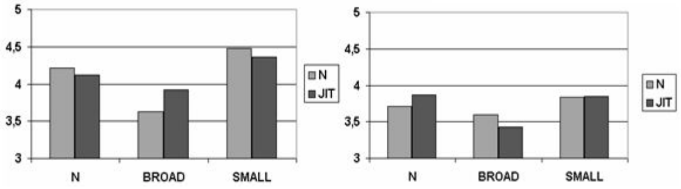

Figure 1: Effects of pitch and jitter for sad (left) and frightened (right) message (N: neutral/original, JIT: jitter)

This result is stronger if you consider the combined effect of jitter-simulation, pitch range and phrase, which was also significant. For the sad message, depicted in Figure 1 left side, the jitter-simulation was considered better when the pitch-range did not suit well. With the frightened message (results depicted in right part of figure 1), the jitter simulation worked better with the appropriate small and normal pitch-range.

6.4. Differences Between Languages

Language in itself did not reach a significant difference in the judgments, but there were significant differences in conjunction with all main inner-subject factors sentence, pitch, duration and jitter. If we compare the results from the different countries, we must of course always keep in mind that the stimuli for each country were not identical. Therefore deviations may be based on other facts than just the cultural difference.

- Maybe the differences are coming from the synthesizer itself. The synthesis is done from diphones of different speakers, and the quality for each language may vary.
- The underlying prosody originates from different speakers and the rules generate different outcomes.
- The translation of the sentences may have resulted in different semantics implying a different kind of appropriate emotion.

Irrespective of the manipulations, the German listeners found the neutral sentence most appropriate displayed, French listeners preferred the boring and sad one whereas the Turkish liked the friendly one best. We cannot be sure if this is primarily an effect of different culture or originates from different semantics of the translations of the original sentence into the target languages. We will confine the further discussion on results that not only differ in the values but are diametrically opposed.

The effects of pitch range for the frightening message (Figure 2) denote that French and Greek listeners prefer a small pitch-range while German and Turkish did not make a distinction. For the neutral sentence contrary to all other countries French listeners prefer a broad pitch-range, while Turkish and Greek listeners distinctively don’t. In contrast to the other countries, the Turkish seem to find a small pitch range acceptable for a friendly expression.

Figure 3 shows that the Greek and Turkish listeners had clearly different opinions regarding whether sad speech should

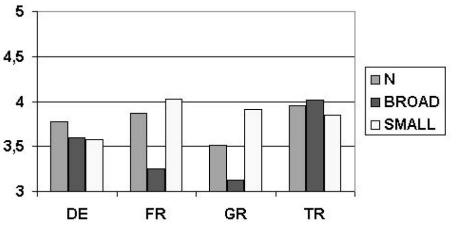

Figure 2: Effect of pitch-range on frightening message depending on language (N: neutral / original pitch range)

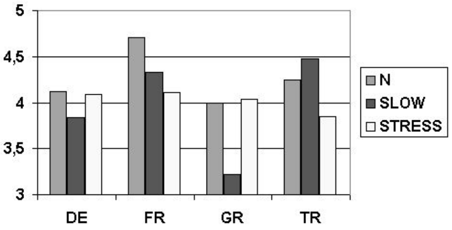

Figure 3: Effect of duration on sad message depending on language (N: neutral / original durations)

be slow or not, whereas the French and the German didn't decide.

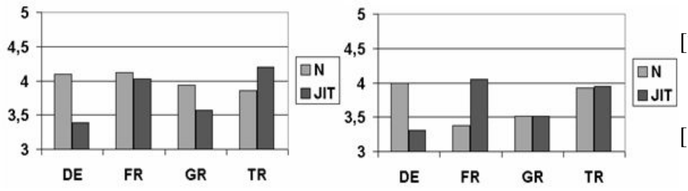

Figure 4: Effects of jitter for threatening (left) and frightening (right) messages (N: neutral, JIT: jitter)

In figure 4 the results for the jitter-simulation for the threatening and the frightening phrase are displayed. While German and French listeners found the jitter-simulation clearly unthreatening, the Turkish thought otherwise. With the frightening phrase the disagreement can be found between German and French; while for the French the frightened impression was enhanced by jitter-simulation as supported by literature (e.g. [14]), for the Germans it was degraded. The Greek and Turkish listeners didn't decide, which in a way can be regarded as a support for jitter, because generally the jitter-simulation clearly lowered the acceptance of the stimuli.

7. Discussion

The results show generally an agreement with the hypotheses derived from the literature and indicate that the listeners, irrespective of the language, interpreted the semantics of the phrases as intended, perhaps with the exception of the phrase meant to be bored.

Nonetheless, there were differences between the different countries, and although we can not be sure that all effects were based on cultural difference alone, we feel that a cross cultural global emotion simulation will not work as expected. These findings also indicate that results based on data-analysis from different cultures can not be applied without reservations.

8. Acknowledgments

This work was partially funded by the EU FP6-IST project HUMAINE (Human-Machine Interaction Network on Emotion).

9. References

[1] Abelin, A.; Allwood, J., 2000. Cross Linguistic Interpretation of Emotional Prosody, *Proc. ISCA Workshop on Speech and Emotion.* Belfast.

[2] Braun, A.; Katerbow, M., 2005. Emotions in Dubbed Speech: An Intercultural Approach with Respect to F0. *Proc Interspeech*. Lisbon.

[3] Burkhardt, F., 2005. Emofilt: the Simulation of Emotional Speech by Prosody-Transformation. *Proc. Interspeech.* Lisbon

[4] Burkhardt, F.; Sendlmeier, W.F., 2000. Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis. *Proc. ISCA Workshop (ITRW) on Speech and Emotion.* Belfast

[5] Cabral, J. P.; Oliveira, L. C., 2005. Pitch-synchronous time-scaling for prosodic and voice quality transformations, *Proc. Interspeech 2005.* Lisbon.

[6] Cahn, J. E., 1989. The affect editor. *Journal of the American Voice I/O Society.* vol. 8, pp. 1-19.

[7] Drioli C.; Tisato G.; Cosi P.; Tesser F., 2003. Emotions and Voice Quality: Experiments with Sinusoidal Modeling. *Proc. of Voqual 2003*

[8] Dutoit, T.; Pagel, V.; Pierret, N.; Bataille, F.; Van der Vreken, O., 1996. The Mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. ICSLP'96* Philadelphia, vol. 3, pp. 1393-1396.

[9] Scherer, K. R.; Banse, R.; Wallbott, H. G., 2000. Emotion Inferences from Vocal Expression Correlate across Languages and Cultures. *Journal of Cross-Cultural Psychology* 2000.

[10] Scherer, K. R., 1986. Vocal Affect Expression: A Review and a Model for Future Research. *Psychological Bulletin* 99(2):143-165.

[11] Schröder, M., 2004. Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. *PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics* Saarland University.

[12] Schröder, M., 2001. Emotional speech synthesis - a review. *roc. Eurospeech 2001* Aalborg, pp. 561-564.

[13] Türk, O.; Schröder, M.; Bozkurt, B.; Arslan, L. M., 2005. Voice quality interpolation for emotional text-to-speech synthesis. *Proc Interspeech 2005* Lisbon.

[14] Williams, C. E.; Stevens, K. N., 1972. Emotions and Speech: Some Acoustical Correlates. *Journal of the Acoustical Society of America (JASA)* 52:1238-1250.