# IRONIC SPEECH – EVALUATING ACOUSTIC CORRELATES BY MEANS OF SPEECH SYNTHESIS

*Felix Burkhardt*[1]*, Alexandra Steinhilber*[2] *and Benjamin Weiss*[3]

[1]*Deutsche Telekom AG,* [2]*Université de Grenoble,* [3]*Technische Universität von Berlin*
*felix.burkhardt@telekom.de*

**Abstract:** We describe an approach to evaluate the acoustic correlates of ironic speech by means of speech synthesis. We manually extracted specific acoustic strategies to generate an ironic expression and synthesized them with a speech synthesizer with the aim to evaluate the perceptual relevance of these strategies. We specifically investigated four distinguished rules: exaggerated articulation, overstressing the main focus-syllable, raising the pitch at the end and speaking with extremely low arousal. The perceptual relevance of the modifications were tested with a web-based listening experiment with 63 native German listeners. All four strategies of prosodically signaling irony by synthesis resulted in a preference over the default version.

## 1 Introduction

Verbal irony is a communicative strategy that occurs when someone says something that is obviously not expressing the real intention or meaning. Often, it is even the opposite is meant. According to the Relevance Theory [1], irony in speech can be considered as an attitude towards one's own statement, consequently modifying its meaning. Irony resembles an indirect expression of critique, and/or humor, and is explicit, as the addressee is supposed to infer from common ground and the specific situation the discrepancy between the speaker's intention and the semantics of the sentence.

Ironic utterances are often marked by a specific prosody (so called "dripping" irony), which is, however, not mandatory (a "dry" version) [2]. It is not clear, whether there is a purely acoustic form of irony, irrespective of the words being uttered. Studies have not been very promising, indicating that prosody alone is insufficient to signal irony [2].

Nevertheless, prosodic signals of irony have been found to often accompany ironic utterances. Single target words exhibit acoustic differences between ironic from literal meaning, for example in female Germans. The ironic versions were lower in fundamental frequency (F0), had longer and hyperarticulated vowels, and more energy [3]. A similar study for French females revealed an expanded F0 range with higher average F0, syllable lengthening, and a raised F0 contour instead of a falling one as discriminating ironic from sincere versions in target words [4]. A subsequent re-synthesis experiment on the same target words confirmed the effects for all three features combined, for lengthening only, and pitch contour only. For English, particularly a lower average F0, but also lower F0 range, longer durations and higher intensity are found in ironic conditions [5, 6]. Directly contrasting adjunct sentences, ironic utterances were slower [7].

There is an assumed incongruence in affect, expressed by the ironic prosody, with the valence of the semantics for dripping irony [8, 9], for example an unhappy tone for a semantically positive statement (of an unpleasant situation). The effects of ironic utterances can even be observed in fMRI data [10]. From our daily experience probably most of us will tend to agree that

this discrepancy between content and prosodic attitude can be easily detected – for the dripping case of intentionally signaling this ironic attitude at least.

Irony detection is obviously an important step in human-machine communication, as it can reverse the meaning of an utterance. One use-case would be to enhance the semantic processing of robots and software agents with speech interfaces, for which not only the recognition, but also synthesis of ironic prosody is required. As previous work, we thus conducted an experiment to investigate verbal irony by distributing a smartphone application that evokes and records user utterances and recognizes ironic speech by detecting discrepancies between the acoustics and sentence semantics. Data was collected by a group of test persons, using the application, and subsequently annotated in order to test and improve the recognizer [11]. This data is now the basis for further investigation.

In the follow up work presented here, we manually extracted specific acoustic strategies to generate an ironic expression and synthesized them with a speech synthesizer with the aim to evaluate the perceptual relevance of these strategies. We specifically investigated four distinguished rules: exaggerated articulation, over-stressing the main focus-syllable, utterance final raising of pitch, and speaking with extremely low arousal.

This article is structured as follows. Firstly we describe the speech synthesizer in section 2. Section 3 reports on the data collection that is the basis for this experiment. We then report on the way we prepared sample sentences in section 4. The next section 5 describes the perception experiment that was used to evaluate the distinguished strategies. Lastly, section 6 discusses first results. We conclude the paper with an overview and some ideas for further studies in section 7.

## 2 Emofilt

Emofilt [12] is a software program intended to simulate emotional arousal with speech synthesis based on the free-for-non-commercial-use MBROLA synthesis engine [13]. It acts as a transformer between the phonetisation and the speech-generation component. Originally developed at the Technical University of Berlin in 1998 it was revived in 2002 as an open-source project and completely rewritten in the Java programming language.

The input format for Emofilt is MBROLA's PHO-format. Each phoneme is represented by one line, consisting of the phoneme's name and its duration (in ms). Optionally following is a set of $F_0$ description tuples consisting of a $F_0$-value (in Hertz) and a time value denoting a percentage of the duration. Here is an example of such a file:

```
_ 50
v 35 0 95  42 95   84 99
O 55 18 99 27 103 36 107 45 111
x 50
@ 30 0 178 16 175 80 160
```

Emofilt's language-dependent modules are controlled by external XML-files and it is as multilingual as MBROLA which currently supports 35 languages.

Emofilt consists of three main interfaces:

- Emofilt-Developer: a graphical editor for emotion-description XML-files with visual and acoustic feedback (see figure 1).
- Emofilt itself, taking the emotion-description files as input to act as a transformer in the MBROLA framework.

- A storyteller interface that can be used to mark phrases in a dialog with colors that correspond to emotional expression [14].

The input format for Emofilt is MBROLA's PHO-format. Each phoneme is represented by one line, consisting of the phoneme's name and its duration (in ms). The valid phoneme-names are declared in the MBROLA-database for a specific voice and must be known by Emofilt.

In a first step each syllable gets assigned a stress-type. Emofilt differentiates three stress-types: unstressed, word-stressed and focus-stressed. As the analysis of stress involves an elaborate syntactic and semantic analysis and this information is not part of the MBROLA PHO-format, Emofilt assigns only focus-stress to the syllables that carry local pitch maxima. However, for research scenarios it is possible to annotate the PHO-files manually with syllable and stress markers, which we did in this experiment.

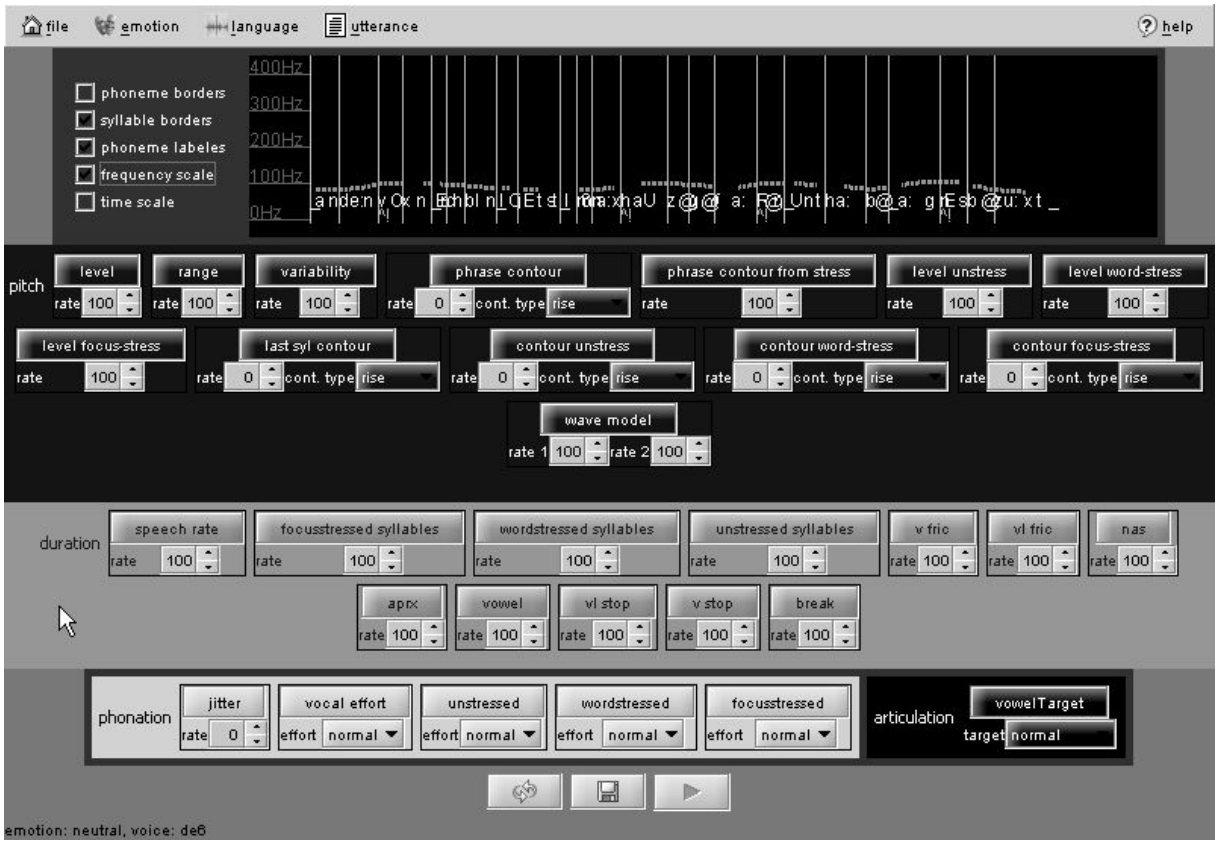


**Figure 1** – Emofilt Developer Graphical User Interface

The emotional simulation is achieved by a set of parameterized rules that describe manipulation of the following aspects of a speech signal:

- Pitch changes, for example: "Model a rising contour for the whole utterance by ordering each syllable pitch contour in a rising manner".
- Duration changes, for example: "Shorten each voiceless fricative by 20%".
- Voice Quality, for example the simulation of jitter by alternating F0 values and support of a multiple-voice-quality database.
- Articulation precision changes by a substitution of centralized and decentralized vowels.


**Figure 2** – Main screen of the irony data collection app, while a result is being shown.

The rules were motivated by descriptions of emotional speech found in the literature [15]. As we naturally can not foresee all modifications that a future researcher might want to apply, we extended Emofilt by an extensible plugin-mechanism that enables users to integrate customized modifications more easily.

## 3 The Ironic Speech Database

To collect data of ironic speech, we created an Android app which is capable of recording audio, streaming it to audEERING's sensAI API service for analysis of emotional parameters, as reported in [11] The app displays the result to the user immediately after processing. For the preliminary irony detection module, a text transcription based on the Google Speech Services API and a ternary text sentiment classification was added to audEERING's sensAI API. Further, for the purpose of collecting a spoken irony database, all requests to the API made through this app were saved on the back end in order to build the database from these recordings.

Figure 2 displays the main screen of the irony data collection app, while it is showing a recognition result. As can be seen, results for the acoustic analysis (activation and valence dimension) are displayed as well as a sentiment result based on the automatically recognized text. An additional anger level is displayed but this is not related to the irony research. If the textual sentiment and acoustic valence diverge, the irony warning button at the top starts to blink.

We conducted a workshop with lay people to gain experience on how the automatic recognition of sentiment will be perceived by users. During the workshop, 12 test users got introduced to the app and tried it out. The testing subjects installed the app on their private Android phones and were instructed to use it for one week. After the try-out period a set of 937 labeled samples

had been collected. This set was then labeled by six listeners for display of anger and irony. The data-set can be requested from the main author for research purposes.

# 4 Sample Generation

In this experiment, we manually extracted specific acoustic strategies to generate an ironic expression and synthesized them with a speech synthesizer with the aim to evaluate the perceptual relevance of these strategies. We specifically investigated four distinguished rules: exaggerated articulation, over-stressing the main focus-syllable, raising the pitch at the end and speaking with extremely low arousal.

The purpose was to generate sentences that potentially would sound ironic. Participants would have to listen to pairs of sentences and decide which one sounds more ironic. For each pair, the text would be the same, but the intonation would be different.

## 4.1 Choosing the sentences to generate

To be more representative, we choose sentences that were obviously ironic according to the text, and others that were not. Moreover, the sentences were pronounced by a man and a woman, for the same purpose. The two sentences who were obviously ironic were "*Ich liebe Regenwetter*" (*I love rain*) and "*Stau, wie toll!*" (*Traffic jam, how great!*). The two sentences who were not it inherently ironic were "*Ich finde das wunderschön*" (*I think it's wonderful!*) and "*Das ist richtig gut*" (*it's really good*).

For each sentence, five groups of emotional expressions were tested:

- Low arousal : the sentence is spoken slowly, the pitch of the sentence is low.
- Focus : the focus phoneme is slow and high.
- Articulation : the sentence is pronounced loud and overarticulated.
- End : the end of the sentence is higher.
- Nothing: no parameter added.

The main stressed syllable of the sentences, i.e. the focus of the sentence, was manually set. For example, in the sentence "it's really good", an accent can be put on the syllable "rea" of "really" to sound more ironic.

Every expression was pair-wise combined for each sentence (for example Low Arousal and Focus for the same file, or only Focus). Like this the test listeners could judge the files by comparison ("which sounds more ironic?") as opposed to a direct judgement.

## 4.2 Resynthesizing sentences

Emofilt usually takes an input text to synthesize speech. It uses MaryTTS, which generates a phonetic file from the text. Emofilt parameters modify the phonetic file, and an audio file can then be generated with Mbrola.

However, if we automatically generate the files with a text as input, the prosody doesn't sound natural. That's why the phonetic files were generated with an audio file by a software named WaveSurfer. One man/one woman pronouncing the four sentences as neutral as possible were recorded. The extraction of the intonation of the audio file was made automatically by the software. The phonemes were manually separated with annotations, to indicate which one there

are (“a”, “e” etc.) and how long they last. The new phonetic files could then be generated and used by Emofilt to create the audio files with all different emotions and sentences.

We then recorded four German sentences from our database, two of them with a strong potential for irony and two with a much lesser potential. An example for the former would be “Ich liebe Regenwetter” (I love rain) and an example for the latter: “Das ist wunderschön” (This is really beautiful).

We recorded these sentences with a male and a female speaker each, spoken with a neutral expression. The prosodic and phonetic descriptions of these recordings were then extracted based on a manual phonetic transcription with a spectrographic analysis software (Wavesurfer).

The resulting PHO-files (input for the Mbrola synthesizer [2]) were assigned a focus syllable manually and then processed by the Emofilt [2] emotional software to apply the above mentioned strategies on the neutral version. We used a male and a female voice for all four sentences and compared each of the four rules plus the neutral version pair-wise, so we ended up with 80 comparisons (5*2 comparisons times two speakers times four sentences).

The following listing displays the modification rules in the Emofilt XML format.

```
<emotion name="Focus" color="ffffff">
        <duration>
        <durationFocusstressedSyls rate="200"/>
        </duration>
        <pitch>
        <levelFocusstress rate="120"/>
        </pitch>
</emotion>
<emotion name="Ende" color="ffffff">
        <pitch>
        <lastSylContour rate="20" type="rise"/>
        </pitch>
</emotion>
<emotion name="LowArousal" color="ffffff">
        <pitch>
        <f0Mean rate="80"/>
        <variability rate="80"/>
        </pitch>
        <duration>
        <speechRate rate="130"/>
        </duration>
</emotion>
<emotion name="Articulation" color="ffffff">
        <articulation>
        <vowelTarget target="overshoot"/>
        </articulation>
        <phonation>
        <vocalEffort effort="loud"/>
        </phonation>
</emotion>
```

## 5 Perception Experiment

The 80 synthesized stimuli were played back to 63 native German listeners (36 female, 27 male, aged 19–52, M=29.9, SD=8.23) using a browser-based Graphical User Interface based on *beaqleJS*[1], each answering the question for all pair-wise comparisons: "which of the two sentences sounds more ironic?". 13 of these 63 participated remotely, whereas the remaining listeners came to an experimental room and used AKG K-601 headphones. Those participants received financial compensation for their effort. The order of the pairs was randomized and the order of the stimuli within each comparison was reversed for half the number of participants.

The preference ratings for each of the ten combinations of irony synthesis strategy were separately tested with binominal tests. These used the number of preferences from all eight conditions of speaker and sentence, thus eight times 63 binary values. Applying Bonferroni-Holm adjustments, there are nine out of ten pairs showing a significant preference. Only the pairwise comparison three does not show a clear preference.

**Table 1** – Results of the binominal tests, Bonferroni-Holm corrected.

| comparison | significance | preferred condition |
|---|---|---|
| 1: Articulation vs. End | $< .0001$ | Articulation |
| 2: Articulation vs. Focus | $< .0001$ | Focus |
| 3: Articulation vs. LowArousal | 0.8237 | n.a. |
| 4: Articulation vs. default | $< .0001$ | Articulation |
| 5: End vs. Focus | $< .0001$ | Focus |
| 6: End vs. LowArousal | $< .0001$ | LowArousal |
| 7: End vs. default | $= .0170$ | End |
| 8: Focus vs. LowArousal | $< .0001$ | Focus |
| 9: Focus vs. default | $< .0001$ | Focus |
| 10: LowArousal vs. defaul | $< .0001$ | LowArousal |

As depicted in Table 1, all conditions with the aim of being ironic are rated significantly more ironic than the default synthesis. The ranking is as follows:

1. Focus
2. Articulation/LowArousal
3. End
4. default

## 6 Discussion

All four strategies of prosodically signaling irony by synthesis resulted in a preference over the default version. The strongest effect was found for the exaggerated sentence stress, synthesized with longer duration and higher pitch. A similar effect might also be created by lowering the pitch of this syllable [3]. While the forced choice paradigm does not tell anything about the naturalness of the prosody accompanying the ironic sentence, it is a first step to use this acoustic signal to be more transparent in artificial agents using ironic utterances for communication. Apart from enhancing the capabilities of recognizing and synthesizing irony in human-computer

[1] http://hsu-ant.github.io/beaqlejs/

interaction, the synthesis strategies evaluated here have to be tested in real communicative situations to be understood by human users.

The text category (inherently ironic text versus positive text) did not make a difference in the listener's judgements, with the exception, that the manipulations with respect to "Focus" and "Low arousal" were perceived as more ironic with the inherently ironic phrases.

## 7 Conclusions and Outlook

In this experiment, we manually extracted specific acoustic strategies to generate an ironic expression and synthesized them with a speech synthesizer with the aim to evaluate the perceptual relevance of these strategies. We specifically investigated four distinguished rules: exaggerated articulation, over-stressing the main focus-syllable, raising the pitch at the end and speaking with extremely low arousal. The test samples were evaluated in an listening experiment and all modifications led to a strengthening of the impression of irony in the utterances.

For further research, these strategies could be tried out with other methods of speech synthesis, like for example non-uniform unit-selection or formant synthesis. Also the data base could be enlarged with more carrier sentences and more modification strategies. Furthermore, a multicultural comparison study would be interesting, for example to investigate in how far verbal irony can be perceived by listeners from different cultures that might not even speak the target language.

## References


[1] WILSON, D. and D. SPERBER: *Explaining irony*. In *Meaning and Relevance*, pp. 123–146. Cambridge University Press, Cambridge, MA, 2012.

[2] BRYANT, G. A. and J. E. FOX TREE: *Is there an ironic tone of voice? Language and Speech*, 48(3), pp. 257–277, 2005.

[3] SCHARRER, L., U. CHRISTMANN, and M. KNOLL: *Voice modulations in German ironic speech. Language and Speech*, 54(4), pp. 435–465, 2011.

[4] GONZÁLEZ-FUENTE, S., P. PILAR, and I. NOVECK: *A fine-grained analysis of the acoustic cues involved in verbal irony recognition in French*. In *Proc. Speech Prosody*, pp. 1–5. 2016.

[5] CHEANG, H. S. and M. D. PELL: *The sound of sarcasm. JSpeech Communication*, 50(5), pp. 366–381, 2008.

[6] ROCKWELL, P.: *Lower, slower, louder: Vocal cues of sarcasm. ournal of Psycholinguistic Research*, 29(5), pp. 483–495, 2000.

[7] BRYANT, G. A.: *Prosodic contrasts in ironic speech. Discourse Processes*, 47(7), pp. 545–566, 2010.

[8] BRYANT, G. A. and J. E. FOX TREE: *Recognizing verbal irony in spontaneous speech. Metaphor and Symbol*, 17(2), pp. 99–117, 2002.

[9] WOODLAND, J. and D. VOYER: *Context and intonation in the perception of sarcasm. Metaphor and Symbol*, 26(3), pp. 227–239, 2011.

[10] MATSUI, T., T. NAKAMURA, A. UTSUMI, A. T. SASAKI, T. KOIKE, Y. YOSHIDA, T. HARADA, H. C. TANABE, and N. SADATO: *The role of prosody and context in sarcasm comprehension: Behavioral and fMRI evidence. Neuropsychologia*, 87, pp. 74–84, 2016.

[11] BURKHARDT, F., B. WEISS, F. EYBEN, J. DENG, and B. SCHULLER: *Detecting vocal irony*. In *Proc. GSCL*, pp. 1–8. 2017.

[12] BURKHARDT, F.: *Emofilt: The simulation of emotional speech by prosody transformation*. In *Proc. Interspeech 2005, Lisbon*. 2005.

[13] DUTOIT, T., V. PAGEL, N. PIERRET, F. BATAILLE, and O. VAN DER VREKEN: *The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. Proc. ICSLP'96, Philadelphia*, 3, pp. 1393–1396, 1996.

[14] BURKHARDT, F.: *An affective spoken story teller*. In *Proceedings of Interspeech*. 2011.

[15] BURKHARDT, F.: *Simulation emotionaler Sprechweise mit Sprachsynthesesystemen*. Shaker, 2000.