An Investigation of Intensity Patterns for German

Oliver Jokisch and Marco Kühne

Laboratory of Acoustics and Speech Communication
Dresden University of Technology, 01062 Dresden, Germany

oliver.jokisch@ias.et.tu-dresden.de

Abstract

The perceived quality of synthetic speech strongly depends on its prosodic naturalness. Concerning the control of duration and fundamental frequency in a speech synthesis system, sophisticated models have been developed during the last decade. Speech intensity modeling is often considered as algorithmically and perceptually less important. Departing from a syllable-based, trainable prosody model the authors tested new factors of influence to improve the predicted intensity contour on phonemic level. Therefore, a German newreader corpus has been analyzed with respect to typical intensity patterns. The f0-intensity interaction has the most significant influence and was perceptually evaluated by 32 listeners ranking 20 different stimuli. Using an elementary, linear intensity model, modified natural speech only slightly degrades about 0.3 at the ITU-T conform MOS scale.

1. Introduction

The intelligibility and the perceived naturalness of synthetic speech strongly depend on the prosodic quality. Recent systems concatenating larger chunks of speech from a database achieve a considerably high quality, as they widely preserve the natural prosodic structure. Because of resource limitations prediction and control of prosodic parameters will keep their importance.

The importance of intensity including its interactions with pitch and timing is widely acknowledged, but only a few institutions devote work to intensity modeling so far, as Monaghan noticed in [1]. Recently, Mixdorf and Jokisch published an integrated, trainable model [2] predicting f0, duration and intensity of syllables. To improve accuracy and perceptual acceptance of the predicted intensities by using this model, the authors interpreted earlier intensity studies, as e.g. [3], [4] or [5], and tried to find similar factors of influence in an available German news corpus.

2. Database and analysis method

2.1. German news corpus

The analyzed data are part of a German corpus compiled by the Institute of Natural Language Processing, University of Stuttgart and consists of 48 minutes of news messages from the radio station Deutschlandfunk, read by a male speaker. The database contains 356 sentences with 5726 words including 13152 syllables. The messages were recorded and partly repeated with an offset of 30 minutes. For the investigation 29362 phonemes were available. Regardless of a few word replacements, selected message texts are identically stored in the database. These messages basically differ in their recording dates. The data analysis considers also these recurrences.

The authors point out that this corpus does not contain spontaneous utterances, since news reading style and individual speaker characteristics are well-defined and reproducible. With regard to the prosodic target model for speech synthesis this reading corpus seems to be appropriate.

2.2. Intensity measures

This study addresses the variations of measurable intensity parameters as the short-time amplitude and the root mean square value of units from the described news corpus which characterizes one possible speaking style for a target synthesizer. Both different intensity measures have been examined (equations 1 and 2). The complete analysis was conducted in the time domain.

\[
\hat{A}_{N} = \max_{k} \left\{ |x(k)| \right\} \quad (1)
\]

\[
RMS_{p} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} x^{2}(k)} \quad (2)
\]

The initial decision has to identify the appropriate linguistic object for the intensity measurement and whether the local intensity contour inside this object is relevant, too. In [2], the authors already measured syllabic RMS values in the news corpus in order to train a neural network model. In this study, the authors focus on the phonemic object to compare the results and to observe local intensity variations inside the syllable. To roughly specify the intensity contour inside a phoneme, all phonemes were separated in different sections. The authors measured following intensity values:

- RMS in the center of the phoneme,
- RMS at the begin of the phoneme,
- RMS at the end of the phoneme,
- RMS over the complete phoneme,
- Short-time amplitude over the complete phoneme.

3. Factors of influence

In [3], the authors found following important factors of influence on the phoneme intensity in the French language:

- Fundamental frequency f0,
- Left and right phoneme context,
- Position in syllable and phrase,
- Duration.

At first, the authors examined similar intensity variations in the German corpus to implement according factors of influence. The correlation between the mentioned factors of influence and the intensity structure was analyzed.
3.1. Intra-individual scattering

To examine the intra-individual scattering, only a single male speaker from the database was available, speaking the same news multiple times per day. Since the news messages were repeated after 30 minutes, identical messages from this speaker could be recorded, i.e. intensity variations of identical messages were gathered as well.

Figure 1 shows different, phoneme-based intensity contours of a typical sentence from the corpus. The different contours represent identical messages spoken at different times.

3.2. Phoneme position

This section discusses the influence of the phoneme position in phrase and syllable. Figure 2 summarizes log-scaled mean values and standard deviations of the syllable components onset, nucleus and coda in the observed corpus.

Without considering other factors of influence, the overall analysis proves that syllables (including their phonemes) have a clearly higher intensity at sentence begin than the corresponding objects at sentence end. Figure 3 shows the comparison of both intensity levels.

3.3. Contextual influence

To visualize contextual influences, the authors exemplarily measured the influence of neighboring phonemes on the intensity of long vowels. Corresponding to [6] the authors compared the intensity in the center of the vowel with the intensity in the right boundary section depending on the right phoneme context. Following phoneme categories for the right context were distinguished:

- Plosive, in particular plosive pause,
- Fricative,
- Long vowel,
- Short vowel,
- Nasal,
- Liquid.

Figure 5 shows the intensity drop at the right phoneme boundary versus the mean intensity in the phoneme center depending on the succeeding phoneme category. For this purpose the RMS value has been selected as intensity measure.

With respect to the influence caused by the left context the data analysis in [6] demonstrates the lower significance of this influence. Also in the current study, the authors observed a lower influence on the intensity at the right boundary.
3.4. Interaction with other prosodic parameters

Finally, the influence of the fundamental frequency $f_0$ has been examined. In [4] and [7], the authors found interesting indications for a strong interaction between intensity and $f_0$. For the Danish language Brondsted was using the RMS contour as an indicator for poorly predicted $f_0$ values. Oppositional contours often (but not necessarily) unmask an incorrect extraction of the $f_0$ contour by the underlying pitch tracking algorithm.

The investigation shows a similar interaction for the analyzed German corpus. Figure 6 visualizes this synchronism between intensity and $f_0$ for an example section of a sentence. Both intensity measures, RMS and short-time amplitude are diagrammed.

The $f_0$-intensity interaction over all phonemes from the database has been calculated by means of the correlation coefficient according to Neyman-Pearson (see table 1).

The scatter plot in figure 7 merges the intensities of all vowels, nasals and liquids. The figure visually suggests a partly proportional interrelation between mean intensity and mean $f_0$ for these phoneme categories. One can obtain similar results if these categories are separately considered.

### Table 1: $f_0$-intensity correlation over all data.

<table>
<thead>
<tr>
<th>Intensity measure</th>
<th>Correlation coefficient $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-time amplitude</td>
<td>0.69</td>
</tr>
<tr>
<td>RMS</td>
<td>0.72</td>
</tr>
</tbody>
</table>

4. Perceptual experiment

32 listeners, 10 females and 22 males (age between 20 and 50 years), evaluated news messages with regards to naturalness and intelligibility. The listeners were subdivided into 16 experts (partly familiar with speech synthesis) and 16 non-professional listeners. Considering the ITU-T conform Mean Opinion Score (MOS), the absolute category rating was conducted at a scale from 1 (“bad”) to 5 (“excellent”). The message length amounted to 6 seconds in average. Both, natural and synthesized speech prompts were presented.
The test set contained five different messages. For each message alternatively following waveform signals have been prepared:

- Original signal (natural voice) without modification (ORG),
- Original signal modified by a linear intensity model (ORGMOD),
- Pure diphone synthesis without any intensity manipulation using duration and f0 models (SYN),
- Diphone synthesis with (mapped) original durations and (mapped) original f0 contour (MAP),
- Synthesis signal as before modified by the mentioned, linear intensity model (MAPMOD).

All messages could be replayed if necessary. The probands had no information about the mentioned signal categories or modification principles.

To test the perceptual relevance of the described factors of influence the authors implemented an elementary intensity modification following a linear approach (equation 3):

$$RMS = a \cdot f_0 + b \quad \text{with} \quad f_0 \geq 60 \text{ Hz}$$ (3)

The coefficients a and b have been extracted from the speech corpus by linear regression. In unvoiced sections the intensity contour was interpolated and smoothed.

The results suggest that an elementary, linear approach in principle allows an intensity control with respect to the interacting f0 parameter. The modification of natural speech with this linear approach degrades the MOS ranking only about 0.3 categories. Nevertheless, the additional intensity modification of widely monotone concatenation units causes quality losses in synthesis. The perceptual evaluation of the modifications reflects at most the same quality as non-modified synthetic speech (group non-experts). This problem is caused by signal modification itself and would probably not occur by using a corpus-based synthesis system. The authors expect further improvements by integrating the other factors of influence and by extending the syllable-oriented neural network model [2] with the results from this study.

The question in dispute, whether an intensity control can significantly improve a concatenative synthesizer, could not be answered in this study. In [8], the authors more precisely examine the practical influence of prosodic components on the overall evaluation of a TTS system.

5. Conclusions

The huge evaluation gap between original and synthesis is in consequence of a direct confrontation without “worst case stimuli” as a bottom reference to provoke discriminative decisions. In synthesis evaluations the used diphone synthesizer achieves a MOS of about 3.0. For the experiment in this study the relations between ORG and ORGMOD as well as between MAP and MAPMOD are relevant.

Table 2: MOS evaluation results (experts and non-experts).

<table>
<thead>
<tr>
<th>Group E&amp;N</th>
<th>MOS</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORG</td>
<td>4.82</td>
<td>0.23</td>
</tr>
<tr>
<td>ORGMOD</td>
<td>4.50</td>
<td>0.31</td>
</tr>
<tr>
<td>MAP</td>
<td>2.48</td>
<td>0.44</td>
</tr>
<tr>
<td>MAPMOD</td>
<td>2.27</td>
<td>0.38</td>
</tr>
<tr>
<td>SYN</td>
<td>1.92</td>
<td>0.44</td>
</tr>
</tbody>
</table>

The question in dispute, whether an intensity control can significantly improve a concatenative synthesizer, could not be answered in this study. In [8], the authors more precisely examine the practical influence of prosodic components on the overall evaluation of a TTS system.

6. References


