CONSISTENCY OF PROSODIC TRANSCRIPTIONS LABELLING EXPERIMENTS WITH TRAINED AND UNTRAINED TRANSCRIBERS

Matthias Reyelt
Institut für Nachrichtentechnik, Technische Universität Braunschweig, Germany
email: m.reyelt@tu-bs.de

ABSTRACT

For the use in the German Project VERBMOBIL a labelling system has been designed, that can be used by different project partners for a variety of purposes. It is based on the ToBI system for English and has some extensions to satisfy the special needs of individual project partners.

As the prosodic transcriptions have to be made by several transcribers having only little training, the achievable consistency was examined in several experiments.

Two main labelling experiments are described here: the first, with fully untrained transcribers in order to obtain a starting value for inter transcriber consistency; the second, after a training phase to examine the improvement in consistency achieved by the training.

INTRODUCTION

Although prosody has been investigated for several decades, the resulting knowledge has rarely found its way into automatic speech recognition. One reason for this might be that the statistical methods like HMM and statistical grammars, that seem to be a current standard in speech recognition, need large amounts of labelled speech data for training to produce reliable results.

However, for the recognition of spontaneous speech also prosodic information is needed. One of the aims of the German project VERBMOBIL is the integration of prosodic information at all levels of the recognition process. The prosodically labelled data needed for training and test are produced centrally for all project partners.

There are several demands made on such a system for prosodic labelling. For a data driven training of speech recognition systems the amount of labelled data must match the number of different labels. If a very detailed inventory is used, a lot of speech material has to be labelled before the data are of any use for automatic speech recognition.

The inventory has to meet the different needs of different users. Such an inventory is always a compromise between machine readability and formal consistency. There are several reasons for introducing this additional tier:

- It is a customer's tier. The information in this tier was needed by partners.
- Together with the break index tier it represents a basic system that can be labelled faster and with less training than a "full" labelling including the tone tier.
- An analysis of the labelled data showed that the syllable durations correspond to the accent type. This tier seems to hold additional information about accents that is not labelled in the tone tier.

The tone tier

In this tier pitch accents, phrase accents and boundary tones are labelled using an inventory similar to ToBI.

The break index tier

This tier, too, is quite similar to the break index tier in ToBI with slight formal changes in the index numbering: intermediate phrase boundary (B2), intonational phrase boundary (B3) and irregular boundary (B9).

EXPERIMENTS

Using this inventory, labelling experiments were carried out. Several subjects made parallel transcriptions of the same material.

In a first experiment [2] [3] five subjects labelled 480 utterances of the PHONDAT92 corpus. The subjects had no experience and only a short introduction to their task. Only the functional tier and a reduced break index tier were used. The transcriptions were based merely on auditory perception, no visual aids such as F0 contour were given.

After this experiment a training programme was developed and in a second labelling experiment the tonal tier
was included as well. For the second experiment 233 utterances from the VERMOMOBIL corpus were used.

**LABELLING ENVIRONMENT**

The labelling was carried out on a workstation using fish, a labelling software based on Tcl/Tk, that is easy configurable and supports the SAM format for labellings.

In the first experiment only the speech signal and the orthographic text was displayed, in the second experiment the pitch contour was added.

**STATISTICAL EVALUATION**

In the first experiment the subjects labelled 480 utterances. The resulting 5520 pairs gave an overall correspondence of 80% for the accents (secondary and main accent) and 94% for phrase boundaries (no further distinctions).

However, this overall correspondence is only a rough evaluation. Additionally the distributions of accent and boundary types are rather unequal and the unaccented syllables make a major contribution to the value.

Thus an independent evaluation value was calculated for each accent/boundary class according to equation 1.

### Equation 1: Calculation of label dependent correspondence

\[
\text{corr}_{i,j,\text{label}} = \frac{n_{\text{correct}_{i,j,\text{label}}}}{n_{\text{lab}_i} + n_{\text{lab}_j}/2}
\]

This leads to the correspondence values shown in Table 1:

<table>
<thead>
<tr>
<th>Table 1: Inter-transcriber correspondence reached by untrained transcribers in the first experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>secondary accent</td>
</tr>
<tr>
<td>main accent</td>
</tr>
<tr>
<td>phrase boundary</td>
</tr>
</tbody>
</table>

The percentages in Table 1 show a satisfying correspondence for main accent and phrase boundary. For secondary accent the correspondence is much lower and shows the transcribers' uncertainty in the decision accent/uncollected.

The overall inter-transcriber correspondence is listed in Table 2.

<table>
<thead>
<tr>
<th>Table 2: Overall correspondence in second experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>functional tier</td>
</tr>
<tr>
<td>break index tier</td>
</tr>
<tr>
<td>tone tier (pitch accent)</td>
</tr>
<tr>
<td>tone tier (boundaries)</td>
</tr>
</tbody>
</table>

Again the correspondences for the individual labels were calculated. Table 3 shows the results for the functional tier and the break index tier. For the tone tier the correspondence varied widely, from maxima of about 56% for L* and L^+H pitch accents down to an absolute minimum of zero for the downstepped L^+H accent (which occurred only four times). For boundary tones the max. correspondence was 55% for the L-L% boundary, the minimum was 33% for the L-H% boundary.

### Table 3: Correspondences for individual labels, second experiment

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>main accent</td>
<td>86%</td>
<td>intermediate phrase boundary</td>
<td>44%</td>
</tr>
<tr>
<td>intermed. phrase boundary</td>
<td>90%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The correspondence values are better than in the first experiment, at least for main accent and intonational phrase boundary. For the secondary accent the correspondence has decreased; the distinction accented/uncollected is still rather uncertain.

**ANALYSIS OF THE TRANSCRIPTIONS**

The statistical evaluation gives an overview over the consistency between the transcribers. However it provides no information about the reasons for the different transcriptions and may even hide errors if they are consistently made by all transcribers.

Additionally a more profound analysis of the transcriptions is necessary in order to examine errors and misinterpretations of the labelling system. Such an analysis showed a variety of reasons for differing transcriptions.

Especially the first experiment revealed that consistency is speaker dependent to a high degree. The quality depends on how familiar the transcriber is with the speaker's dialect. Besides, the label inventory and the training do not (yet) cover all German dialects and speaking styles.

Different transcriptions are also caused for several other reasons. Firstly, the categorial boundaries between the labels (e.g. H* and L+H*) are not always clearly distinguishable. Secondly, misinterpretations of the pitch contour lead to erroneous transcriptions. Thirdly, the usage of particular labels was misunderstood by the transcribers.

In an additional training (in particular using erroneous utterances) the number of labelling mistakes can surely be reduced. However, a regular consistency check seems to remain necessary.

**OUTLOOK**

Although these experiments are preliminary, they provided useful insights into practical problems of prosodic labelling. As a result, the training programme has been extended to include the difficult cases.

Moreover the labelling environment has been extended by providing means for the transcribers to mark their uncertainties and to add comments on their transcriptions.

The current database consists of approximately one hour of labelled speech that has already successfully been used by several project partners.

**REFERENCES**


