Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Consistency of Prosodic Transcriptions: Labelling Experiments with trained and untrained transcribers

Matthias Reyelt

TU Braunschweig

Vm

Report 155
August 12, 1996

August 12, 1996

Matthias Reyelt

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
38092 Braunschweig

Tel.: (0531) 391 - 2476
Fax: (0531) 391 - 8218

Gehört zum Antragsabschnitt: 14.6 Prosodische Etikettierung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N 0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Contents

1 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 [5] 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 people.

It is of great importance that the labelled data be ready for automatic analysis. Machine readability and formal consistency are indispensable. A lot of work in the SAM-Project was devoted to the development of label inventories and standard labelfile formats that also include prosodic information [1].

On the other hand there are requirements from the labelling point of view. Larger amounts of data have to be labelled by different transcribers after only a short training phase. Although some subjective variation might be inevitable, the inter–transcriber consistency has to minimized by carrying out several measures:

The label inventory has to be transparent, i.e. it has to match to the perception of transcribers who do not have a profound knowledge of prosodic theory.

A permanent evaluation is necessary to keep track of weaknesses of the system.

A system for labelling large corpora is the ToBI system for English prosody [4]. This system was developped in accordance with the above criteria and has become (or is on its way to becoming) a standard system for transcribing prosody. The labelling system described here is an adaptation of this system for German prosody.

2 Description of the labelling system

The labelling system used in these experiments is divided into three tiers, which are partially similar to the ToBI-system:

2.1 The functional tier

In this tier a more "functional" prosodic labelling is performed. The tier is not part of the original ToBI and is therefore described in detail:

One part of this tier is the labelling of *sentence modality*. This might not be part of a core prosodic analysis but is clearly suprasegmental and is needed by several project partners.

The other part is the basic labelling of *accented words* based on auditive impression. There are three different accent types: *secondary accent*, *main accent* and *emphatic/contrastive accent*. In each intonational phrase the most prominent word obtains the *main accent*[1]. Although this is of course not a focus analysis, it offers some information about the focal structure of the utterance.

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.

2.2 The tone tier

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

2.3 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).

[1] This is not a strict rule; where appropriate, there can be more than one *main accent* per phrase. The *main accent* can be replaced by an *emphatic accent*.

3 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[2]. 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 auditive perception, no visual aids such as F_0-contour were given.

After this experiment a training programme was developped and in a second labelling experiment the tonal tier was included as well[3]. For the second experiment 233 utterances from the VERBMOBIL corpus[4] were used.

4 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 labelfiles.

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

5 Statistic evaluation

In the first experiment the subjects labelled 480 utterances. The resulting 5520 pairs gave an overall correspondence[5] 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 $corr_{1,2,label}$. $n_{corr(1,2),label}$ *is the number of correct pairs for a particular label.* $n_{1,label}$ *and* $n_{2,label}$ *are the total*

[2]The PHONDAT92 corpus consists of single read utterances from a travel inquiry scenario.

[3]Unfortunately only two of the five subjects remained from the first experiment (it seems indeed that prosodic labelling is not that much fun for most people, why?), so the results of this experiment remain preliminary.

[4]The VERBMOBIL corpus consists of spontaneous negotiation dialogues.

[5]This correspondence is calculated according to the ToBI system, see [4]

numbers of this label occurring in each of the transcriptions

$$corr_{1,2,label} = \frac{n_{corr(1,2),label}}{(n_{1,label} + n_{2,label})/2} \tag{1}$$

This leads to the correspondence values shown in Table 1:

Table 1: Inter–transcriber correspondence reached by untrained transcribers in the first experiment

secondary accent	40 %
main accent	72 %
phrase boundary	76 %

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 *accented/unaccented.*

In the second experiment the subjects had a training phase with a number of selected utterances to introduce the label inventory and then a nine–dialogue experience. For the evaluation five different dialogues were chosen, consisting of 233 utterances (2907 pairs). The overall inter–transcriber correspondence is listed in Table 2.

Table 2: Overall correspondence in second experiment

functional tier	91 %
break index tier	94 %
tone tier (pitch acc.)	85 %
tone tier (boundaries)	88 %

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 H* 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 75% for the L-L% boundary, the minimum was 35% for the L-H% boundary.

Table 3: correspondences for individual labels, second experiment

secondary accent	32 %
main accent	86 %
intermed. phrase bound.	44 %
intonational phrase bound.	90 %

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/unaccented is still rather uncertain.

6 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.

7 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 approx. one hour of labelled speech that has already successfully been used by several project partners.

References

[1] *User guide to ETR tools.* SAM-UCL-G007, pp. 15–19, 1992.

[2] Matthias Reyelt. Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German. In *Working papers*, volume 41, pages 238–241, Lund University, 1993. Dept. of Linguistics.

[3] Matthias Reyelt. Untersuchungen zur Konsistenz prosodischer Etikettierungen. In H. Trost, editor, *KONVENS 94*, pages 290–299, Berlin, 1994. Springer.

[4] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. Tobi: A standard for labeling english prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 867–870, 1992.

[5] W. Wahlster. Verbmobil: Translation of face-to-face dialogs. In *Proceedings Eurospeech 93*, 1993.