Measuring Pitch Range

Hanny den Ouden\textsuperscript{1,2} & Jacques Terken\textsuperscript{2} 
\textsuperscript{1}Katholieke Universiteit Brabant Tilburg, \textsuperscript{2}IPO, Center for User-System Interaction, Technische Universiteit Eindhoven 
\textsuperscript{j.n.d.ouden@tue.nl}

Abstract

The literature offers at least two methods to annotators for characterizing the pitch range of a prosodic phrase. One method, included in the ToBI framework, is in terms of the distance between the F0 maximum of the phrase (HiF0) and the speaker’s utterance-final pitch (LoF0). The other method, proposed by Ladd and by ‘t Hart and colleagues, is in terms of the distance between pitch peaks and pitch valleys in the prosodic phrase. In this paper we address two questions. The first question concerns the reliability of the different methods. Five experienced phoneticians applied both methods on a set of forty utterances taken from read aloud text. We found that reliability was higher for HiF0 than for distances between pitch peaks and valleys. The second question is whether variation that is not captured by the ToBI approach, i.e., variation of the baseline, does actually occur in pitch contours. We calculated correlations between values for the different model parameters in the set of forty phrases. The pattern of correlations suggests that the HiF0 approach captures all the variation relevant to measuring pitch range that occurs in our small corpus. We conclude that the HiF0 method is methodologically more adequate, and at the same time sufficient to represent pitch range variation adequately for read aloud text.

1. Introduction

In this paper we address issues concerning the reliability of methods for measuring pitch range and the adequacy of the underlying models proposed in the literature. The basic concepts are illustrated in figure 1, showing an imaginary pitch contour consisting of a sequence of pitch rises and falls. In Germanic languages such as English and Dutch, each combination of pitch rise and fall in Figure 1 would be associated with an accented syllable. The amplitude of the pitch movements may vary, under the conscious control of the speaker, and this aspect of intonation is referred to by the term “pitch range”. It is usually assumed that the intonation phrase is the domain of pitch range variation, so that it need not be measured for each individual pitch movement. That is, for the phrase shown in Figure 1 only one range value needs to be specified instead of five. It has been shown that variation in pitch range has communicative value. For instance, variation in pitch range has been observed in relation to text structure, such that phrases in the beginning of paragraphs have a relatively wide pitch range, whereas phrases near the end of the pitch range have a relatively small pitch range \cite{1}\cite{2}\cite{3}. Similar variation in pitch range has been shown for tone languages as well, e.g. \cite{4}. Different models of intonation contain different accounts of pitch range and pitch range variation. One class of models is based on work by Liberman and Pierrehumbert \cite{5}. This conception of pitch range is also included in the influential ToBI annotation scheme \cite{6}. In this approach, it is assumed that speakers vary the pitch range by scaling the pitch maxima. Accordingly, pitch range is represented in terms of the distance between the highest pitch of the phrase under consideration (HiF0) and a reference line, defined by the lowest pitch of the speaker in the modal register (LoF0). Typically, LoF0 is more or less constant, and occurs at the end of utterances in read-aloud speech and at the end of topical units in spontaneous speech. In fact, the occurrence of a LoF0 is a strong cue for the presence of an utterance end in spontaneous speech \cite{7}. Based on findings from experiments in which speakers were instructed to vary the pitch range explicitly \cite{5}, it is assumed that the baseline is more or less fixed and not affected by manipulations of the pitch range.

![Figure 1. Imaginary, stylized pitch contour illustrating the main concepts. TopBegin and BaseBegin give the height of the Topline and Baseline, respectively, at the beginning of the pitch contour. TopEnd and BaseEnd give the height of the Topline and Baseline, respectively, at the end of the pitch contour.](image-url)

A different class of models can be linked to the IPO approach to intonation \cite{8}. In contrast with the ToBI approach it assumes that speakers can manipulate pitch range by varying both the baseline and the topline, and – moreover – that the baseline and topline can be manipulated independently. If so, it means that pitch range cannot be expressed only in terms of the distance between the HiF0 and the reference line, because this would not include the variation of the baseline. Instead, these models express pitch range in terms of the distance between the topline and the baseline. Representants of this type of model are \cite{9}\cite{10}\cite{11}. There are four kinds of evidence supporting this type of model. In the first place, measurements of pitch range at different positions in text show that the baseline is not constant, and perception experiments with synthesized contours show that modeling this kind of variation gives good acceptability results \cite{3}. Secondly, varying the baseline while
keeping the frequency of the peaks constant will affect the perceived prominence associated with the peaks [11][12]. In the third place, measurements on emotional speech show that different emotions may be associated with peaks in the same frequency range while differing with respect to the valleys [13]. In the fourth place, varying the slope of the baseline influences the decisions of listeners as to whether an ambiguous utterance is a question or a statement [14]. All these findings cannot be accommodated easily within the ToBI framework.

1.1. Measuring pitch range
When we need to measure pitch range to investigate particular questions, a decision must be made as to what method to apply. The choice of a method needs to be based both on considerations of adequacy and reliability. The ToBI method appears to be quite reliable in comparison with the IPO-based approach. Notions such as “the highest accent peak in the phrase” are quite unambiguous, whereas fitting a line through a number of peaks or valleys leaves quite a lot of room for individual annotators and therefore replicability will be difficult to achieve. For instance, one may ask whether the topline drawn in Figure 1 is an adequate trendline for the peaks, and it is not difficult to see that different annotators may come up with widely diverging toplines (see also the debate raised by [15]). Furthermore, one may wonder how to measure a topline or a baseline in a phrase that has only one valley or one peak. Of course, methods can be envisaged to obtain trendlines automatically. While this will increase the reliability, the adequacy of such automatic methods is a much harder problem.

The ToBI approach suffers from similar drawbacks. Most notably, the height of a single peak is subject to multiple influences. It may be a direct reflection of the intended pitch range, e.g. in response to text segmentation. At the same time, it may also express a particular local prominence (emphasis). Finally, the height of a particular peak is also determined by low-level factors such as vowel-intrinsic pitch: peaks associated with the vowel /a/ may be quite a lot higher than peaks associated with the vowel /æ/, and the effect of vowel-intrinsic pitch is strongest in accented syllables where it may induce a difference up to 3 Semitones [6].

Thus, we need to address questions both with respect to the reliability and the adequacy of the different approaches to measuring pitch range. Below, we will first deal with the reliability of measurements across different analysts, and then we will deal with the issue of adequacy.

2. Methods

2.1. Materials
Forty speech utterances were selected from four spoken texts that had been broadcast on Dutch radio. Each text had been read aloud by a different professional speaker. Two texts were telephone news broadcasts, two were commentaries on the news. Ten utterances were selected from each text. These utterances consisted of one intonation phrase and they did not contain end risings, since it was assumed that the presence of end risings might confuse subjects. Half of the utterances were longer than 5 seconds (mean: 6.11 sec.; sd 1.03 sec.); half of the utterances were shorter than 5 seconds (mean: 3.82 sec.; sd .64 sec.). Another set of four utterances was selected, one of each text, as practice materials and to make the judges familiar with the voices of the speakers. The results of these four utterances were left out of the analyses.

2.2. Judges
Five researchers with expertise in prosody served as analysts. Four of them were members of the research programme for Spoken Language Interfaces of IPO at Technical University Eindhoven; one was affiliated with the Phonetic Laboratory of Leiden University. The judges were familiar with the speech processing program they had to use.

2.3. Instruction
Judges performed the analysis individually with the speech processing program Gipos, in which functionality was included to draw and adjust trendlines on the screen by means of cursor clicks. The judges performed the task on their own computer without restrictions of time. They were given instructions as to how to fit and adjust trendlines, and were told to do so on the basis of visual and auditory inspection. After loading a speech file, they first fitted a baseline for the utterance. The line they edited was visible on the screen and could be modified until they were satisfied. Second, after the screen was cleared, the judges fitted a topline of the utterance in the same way as the baseline. Third, they indicated the F0 maximum of the utterance defined as the highest F0 value associated with a pitch accent, ignoring consonantal perturbations and similar micro-prosodic influences. LoF0 was not requested, as we assume that this parameter can be represented by a single constant for the speaker. Frequencies and time stamps for beginnings and endings of the topline and baseline and the HiF0 were automatically put in a text file.

2.4. Data processing
The agreement between the judges is determined with Cronbach’s alpha. By that the range of correlations between the judges is taken into consideration. Procedures for assessing the adequacy of the different models will be described in the results section.

3. Results and Discussion

3.1. Agreement between the judges
Table 1 shows the homogeneity of the pitch characterizations by the five judges and the range of the correlations between the judges with regard to the HiF0 and the beginnings and endings of both lines.

In all cases Cronbach’s α is above .90. By convention values above .70 are called ‘adequate’ and values above .80 ‘good’. A Cronbach’s α above .90 (especially with only five judges) has to be considered as ‘excellent’ [16]. This result gives reason to consider the judgment of the group as a good pitch range characterization. However, the fact that they were homogeneous as a group does not mean that the judges agreed all the time. This was shown by the range of the correlations between pairs of judges. With regard to the characterization of the beginnings and endings of the lines the judges had correlations of at least .59 and at most .87; with regard to the
characterization of the HiF0 the range between the judges varied less (.86-.98). The fact that evidently the judges were no replications of each other implicate that for a reliable characterization of pitch range, the judgments of a group of judges should be representative, but not the individual judgments.

3.2. Adequacy of the models

As explained in the Introduction, the class of models expressing pitch range variation in terms of the distance between pitch peaks and valleys allows independent variation of the upper and lower bound of the pitch contour. This independent variation cannot be captured by the ToBI approach, since it represents the lower bound by a constant and only the upper bound is represented by a variable, HiF0.

In order to determine whether the ToBI approach misses relevant variation, we applied the following approach. For each phrase we calculated mean values for HiF0, TopBegin, TopEnd, BaseBegin and BaseEnd based on the values given by the five phoneticians. In order to eliminate differences in pitch range between speakers, for each speaker parameter values were converted to z-scores relative to his/her average pitch range, i.e., the average of the values for TopBegin, TopEnd, BaseBegin and BaseEnd for all ten phrases. Next, we calculated correlations between sets of z-score parameter values for the whole set of forty utterances to find out whether the value for a particular parameter could be predicted from the value of another parameter. The reasoning is that if the correlation between two parameters is high, the two parameters do not exhibit independent variation. In our analysis, we focused on the correlations between HiF0 and TopBegin, between HiF0 and BaseBegin, and between TopBegin and BaseBegin, because these are precisely the places where the IPO model allows independent variation that cannot be captured in the ToBI approach.

Correlations were .65 between HiF0 and TopBegin, .45 between HiF0 and BaseBegin, and .53 between TopBegin and BaseBegin. These values suggest that HiF0 has predictive value for TopBegin, more so than for BaseBegin. Likewise, BaseBegin cannot be predicted accurately from knowledge of the value of TopBegin. Although this appears to support the claim of the IPO model that BaseBegin and TopBegin may vary independently, inspection of Figure 2 suggests that the low correlation is rather due to the lack of variation for BaseBegin. As can be seen in Figure 2, the range of variation for BaseBegin is quite small, suggesting that BaseBegin can be represented approximately by a constant. Thus, even though perceptual experiments have shown that variation in BaseBegin influences the perception of pitch range, in practice speakers do not seem to exploit this opportunity, at least not in the context of reading aloud informative texts.

Figure 3 illustrates the correlation between HiF0 and TopBegin. As can be seen, as HiF0 goes up, so does TopBegin. Again, in principle the two might vary independently, because HiF0 will also reflect variation of local emphasis on individual words. The most likely explanation for the correlation between HiF0 and TopBegin is that our judges always include the peak associated with HiF0 in their estimate of the topline, irrespective of variations in local emphasis.

Table 1. Per measuring point the homogeneity of the pitch characterizations and the range of the correlations.

<table>
<thead>
<tr>
<th></th>
<th>Cronbach’s α</th>
<th>Range of the correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiF0</td>
<td>.98</td>
<td>.86 - .98</td>
</tr>
<tr>
<td>Topline Begin</td>
<td>.96</td>
<td>.76 - .86</td>
</tr>
<tr>
<td>Topline End</td>
<td>.92</td>
<td>.59 - .78</td>
</tr>
<tr>
<td>Baseline Begin</td>
<td>.95</td>
<td>.79 - .86</td>
</tr>
<tr>
<td>Baseline End</td>
<td>.95</td>
<td>.74 - .87</td>
</tr>
</tbody>
</table>

Figure 2. Relation between z-scores for mean HiF0 and mean BaseBegin for forty phrases (r=.45)

Figure 3. Relation between HiF0 and TopBegin.
4. Conclusion

In this paper we have addressed methodological issues concerning the measurement of pitch range. Several models were considered for representing pitch range, each with their own theoretical assumptions. The ToBI approach is rather straightforward, but may not capture all the variation that has been observed for instance in emotional speech and that has been shown to be perceptually relevant, so that a more sophisticated model might be needed. The problem arises when one starts to apply such a more sophisticated model, including trends like pitch peaks and valleys, to actual prosodic phrases. Automatic procedures for deriving values for the model parameters are hampered by the fact that it needs linguistic intelligence to decide which valleys and peaks one has to take into consideration, and how they should be weighted. We asked two questions. One question concerned the reliability of obtaining parameters for individual phrases. We found that there were considerable differences between parameter values provided by individual analysts, and that reliability was different for different model parameters: reliability was higher for HiF0, and lower for TopBegin, BaseBegin, TopEnd and BaseEnd. Inter-subject reliability was quite good in the sense that high values for one analyst would normally correspond to high values for other analysts as well. Thus, values obtained from individual analysts involve a certain amount of arbitrariness or subjectivity and are no solid basis for modeling trends in pitch range for prosodic phrases. The arbitrariness can be reduced by taking means across subjects. Doing so, we addressed the question of what kind of model would be appropriate to represent trends in pitch range. Basically, the question is whether the ToBI approach captures all the relevant information or whether the more sophisticated Topline-Baseline approach such as advocated in the IPO approach would be needed. The pattern of correlations between different parameters suggests that the ToBI approach seems to capture all the pitch variation that is relevant to pitch range variation, at least for read-aloud text. At the same time, it has the merit of being very reliable. Thus, when measuring pitch range for individual prosodic phrases in read-aloud text, expressing pitch range as the distance between the HiF0 of the phrase and the speaker’s LoF0 is both theoretically and methodologically adequate. Of course, the objections to this method remain. In particular, it is vulnerable for effects of vowel-intrinsc pitch. Thus, conclusions concerning trends for pitch range should be based on quite large amounts of data to eliminate such effects.

5. Acknowledgment

Our thanks go to Carel van Wijk for his helpful comments. The research reported in this paper was financially supported by the Cooperation Unit of Brabant Universities (SOBU).

6. References


