FORMAL AND FUNCTIONAL EVALUATION OF A MELODIC MODEL FOR STANDARD INDONESIAN

Ewald F. Ebing, Vincent J. van Heuven1 & Cecilia Odé

Dept. Languages and Cultures of Southeast Asia and Oceania, Leiden University
1Dept. Linguistics/Phonetics Laboratory, Leiden University

ABSTRACT

A model of Standard Indonesian intonation was perceptually evaluated using an improved testing methodology and listener selection. In a second experiment the focus and boundary marking functions of Indonesian intonation were investigated.

1. INTRODUCTION

A model for Standard Indonesian intonation has been developed following an analysis by synthesis methodology [1,2]. Successive versions of the model were perceptually evaluated by having native Indonesian listeners rate melodic versions of utterances (human originals versus model-generated contours), as well as a priori less adequate melodies, e.g. time-shifted or Dutch contours) along a 10-point scale of formal melodic adequacy [3]. Listeners proved very insensitive to the melodic differences among the versions, so that we decided to re-run the evaluation with (hopefully) improved materials and more carefully selected listeners (section 2). It is difficult in Indonesian to distinguish between the accent-lending and boundary marking function of certain pitch movements [4]. In section 3, therefore, we examine how successfully Indonesian listeners can disambiguate arithmetic expressions with ambiguous focus distribution and internal bracketing.

2. FORMAL EVALUATION

Stimuli were taken from our corpus of quasi-spontaneous monologue by an educated speaker of Indonesian from Riau (East Sumatra) also used in our earlier experiments [2,3]. The stimuli comprised two tokens of the eight perceptually relevant pitch configurations found in our previous experiments. Four melodic versions of each configuration were produced by manipulating $F_0$ in the resynthesis (for procedural details see [3,6]):

a. Close-copy stylizations (COPY) of the human originals; these should receive the highest ratings.

b. Standardized versions (STAN), i.e., generated according to our model; these should be (almost) as acceptable as COPY.

c. Dutch-based versions (DUTCH), generated according to the Dutch intonation grammar [1,3]; these versions should be rated as less acceptable than a or b.

d. Mirrored versions (MIRROR). Close-copies were mirrored along the frequency axis: rises became falls and vice versa; these versions should receive low ratings (as c).

The target configurations were now presented in their original contexts (rather than in isolation). To direct the listeners’ attention to the relevant pitch configuration, the resynthesized context, but not the target configuration, was voiceless (whispered) throughout. This resulted in 64 stimulus types, each presented twice, yielding 128 judgments per listener.

The experiment was run at Universitas Islam Riau in Pekanbaru with 25 university students. Seventeen spoke Riau Malay as their first language, others had a different mother tongue, e.g. Minangkabau. Listeners rated each utterance along a 10-point scale of melodic adequacy (1: extremely poor; 10: excellent).

The results are summarized in Figure 1. The ordering of the acceptability ratings for the entire group of listeners is as predicted. No difference was found between COPY and STAN, t(783)=1.49, ins., nor between STAN and DUTCH, t(777)=1.4, ins. However, the COPY versions were rated as significantly better than the DUTCH-versions, t(779)=3.12, p<.01. The MIRROR versions were rated as poorer than all other versions. Unexpectedly, STAN and DUTCH versions still do not differ significantly.

3. ACCENTS AND BOUNDARIES

The aim of our second experiment was to find out to what extent accentuation and boundary marking can be (independently) expressed by means of the pitch movements in our model.

Focus distribution was manipulated by applying metalinguistic contrasts [5,6]. In the same set of test utterances, we also varied the position of a prosodic boundary by forcing the speaker to disambiguate a potentially ambiguous arithmetic expression (cf. e.g. [8]).

A single male native speaker of Indonesian produced eight versions of the same word sequence ‘dua kali tiga tambah lima’, orthogonally varying the position of the phrase boundary: 2x(3+5) versus (2x3)+5, and focus structure: (1) narrow focus on the first numeral (2) narrow focus on the second numeral (3) narrow focus on the third numeral. Each sentence was prompted by a question sentence to provide a context where one word was placed in focus. By manipulating $F_0$, model-generated contours were made for each realization.

The 25 subjects mentioned above indicated where they thought the speaker had intended the internal bracket of the expression to be, and - in a second part - which one of the three numerals in each phrase carried the strongest accent.

Table 1 specifies the percentage of accent responses for each of the three relevant numerals broken down by intended focus condition, and by intended phrase boundary position, first for the...
Table I. Perceived accents (%) for focus on 1st, 2nd and 3rd numeral, broken down by boundary position: (A) human originals, (B) model-generated contours.

<table>
<thead>
<tr>
<th>Boundaries</th>
<th>Human originals</th>
<th>Model-generated contours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>49</td>
<td>39</td>
</tr>
<tr>
<td>2nd</td>
<td>37</td>
<td>27</td>
</tr>
<tr>
<td>3rd</td>
<td>24</td>
<td>21</td>
</tr>
</tbody>
</table>

In the human originals, accents on the first and second numerals are mostly correctly perceived, although the percentages are lower than we expected, and quite probably lower than what would be obtained with speakers and listeners of English or Dutch. Perception of an accent on the third numeral is strongly disfavoured. Crucially, there is a clear effect of the position of the internal boundary on accent perception: chances of perceiving an accent increase immediately before a phrase boundary. This effect is stronger when focus is on the first syllable than on the second.

For the model-generated contours, the same effects and interactions exist but in a weaker form. When the boundary is after the first numeral, the majority of accents is perceived on the syllables where they were generated, for all three positions: bias disfavouring the third numeral has disappeared. When the boundary is after the second numeral, some bias against perceiving accent on the third numeral remains, but it is clearly weaker than in the human originals. Apparently, our human speaker pronounced very clear accent-lending pitch movements on the first and second, but not on the third numeral. Our model-generated accents were identical for each numeral position, i.e., smaller than the human accents on the first two numerals, but larger than the human accent on the third numeral.

Table II. Correctly perceived phrase boundaries (%) broken down by intended boundary position and focus distribution: (A) human originals, (B) model-generated contours.

<table>
<thead>
<tr>
<th>Boundaries</th>
<th>Human originals</th>
<th>Model-generated contours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>47</td>
<td>42</td>
</tr>
<tr>
<td>2nd</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>3rd</td>
<td>18</td>
<td>20</td>
</tr>
</tbody>
</table>

There is a very strong effect, both for human and for model-generated contours, for more (twice as many) boundaries to be perceived after the second numeral than after the first. It is unclear at this time to what extent this is a stimulus effect. A stimulus analysis (not presented) shows clear differences in duration structure as a function of intended boundary position, but the duration effects are in fact stronger for the first numeral than for the second. Therefore, it seems that the effect is due to linguistic expectancy. There is a smaller effect, both in human and in model contours, to perceive