Phonetic or phonological contrasts in Dutch boundary tones?

Vincent J. van Heuven and Robert S. Kirsner
University of Leiden / University of California at Los Angeles

1. Introduction

Linguistic categorisation of sound. A basic problem of linguistic phonetics is to explain how the infinite variety of speech sounds in actual utterances can be described with finite means, such that they can be dealt with in the grammar, i.e. phonology, of a language. The crucial concept that was developed to cope with this reduction problem is the sound category, or — when applied to the description of segmental phenomena — the phoneme.

Categorisation of sounds may proceed along several possible lines. First, many differences between sounds are simply too small to be heard at all: these are subliminal. The scientific discipline of psycho-acoustics provides a huge literature on precisely what differences between sounds can and cannot be heard with the naked ear. Moreover, research has shown that our hearing mechanism has developed specific sensitivities to certain differences between sounds and is relatively deaf to others. An important concept in this context is the notion of categorical perception. This notion is best explained procedurally in terms of a laboratory experiment.

Imagine a minimal word pair such as English *back* ~ *pack*. The onset of voicing in *back* coincides with the plosive release, whilst the voice onset in *pack* does not start until some 50 ms after the release. We create a series of exemplars by interpolating the voice onset time of a prototypical *back* (0-ms delay) and that of a prototypical *pack* (70-ms delay) in steps of 10 ms, yielding an 8-step continuum ranging over 0, 10, 20, 30, 40, 50, 60, and 70 ms. These eight exemplars are played in random order to English listeners for identification as either *back* or *pack*. The 0-ms voice delay token will yield exclusively *back*-responses (0% *pack*); the 70-ms token will have 100% *pack*-responses. But what results will be

Linguistics in the Netherlands 2004 21 (2004), **102–113**.
ISSN 0929–7332 / E-ISSN 1569–9919

obtained for the intermediate exemplars? If the 10-ms changes in voice delay are perceived *continuously*, one would predict a constant, gradual increase in %-*pack* responses for each 10-ms increment in the delay, i.e., the psychometric function (the line that captures the stimulus-response relationship) is essentially a straight line (open symbols in Figure 1B). The typical outcome of such experiments, however, is non-continuous. For the first part of the continuum all exemplars are perceived as *back*-tokens, the rightmost exemplars are near-unanimously perceived as *pack*. Only for exemplars in the middle of the continuum do we observe uncertainty: here the responses are ambiguous between *back* and *pack*. The psychometric function for this so-called categorical perception is sigmoid, i.e., has the shape of an S (big solid symbols in Figure 1B). In the idealized case of perfect categorical perception we would, in fact, expect a step-function jumping abruptly from 0 to 100% *pack*-responses somewhere along the continuum (thin black line with small solid symbols in Figure 1B).

The *category boundary* (at 35-ms VOT in Figure 1B) is defined as the point along the stimulus axis where the responses are completely ambiguous, i.e., 50–50%. For a well-defined *cross-over* there should be a point along the stimulus axis where 75% of the responses agree on one category, and a second point where there is 75%-agreement on the other. The *uncertainty margin* is the distance along the stimulus axis between the two 75%-points.

Although a pronounced sigmoid function is a clear sign of categorical perception, researchers have always been reluctant to consider it definitive proof. Listeners, when forced to, tend to split any continuum down the middle. Therefore, two conditions should be met: (i) identification should show a clear sigmoid, and (ii) discrimination should show a local peak for stimuli straddling the category boundary.

The discrimination function is determined in a separate experiment in which either (i) identical or (ii) adjacent tokens along the stimulus continuum are presented pair-wise. Listeners then decide whether the two tokens are 'same' or 'different'. Two kinds of error may occur: a physically different pair may be heard as 'same', and a pair of identical tokens may be called 'different'. The results of a discrimination task are best expressed as the percentage of correct decisions obtained for a 'different' stimulus pair minus the percentage of errors for 'same' pairs constructed from these stimuli (the latter percentage is often called the response bias). In the case of true categorical perception the discrimination scores show a pronounced peak for the stimulus pair straddling the category boundary, whilst all other pairs are discriminated at or only little above chance level (see panel A in Figure 1). Physically different sounds that fall in the same perceptual category, are hard to discriminate. In the case of continuous perception, there is no local peak in the discrimination function.

Categorical nature of intonational contrasts? By intonation we mean the pattern of

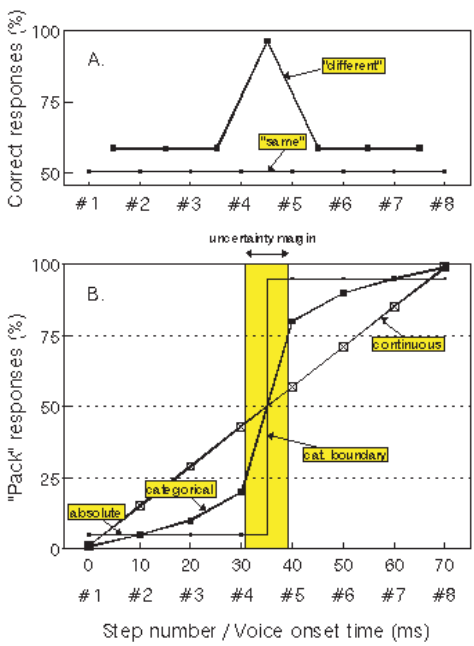

Figure 1. Panel A. Hypothetical discrimination function for physically same and different pairs of stimuli (one-step difference) reflecting categorical perception. Panel B. Illustration of continuous (open squares) versus categorical (big solid squares) perception in the identification and discrimination paradigm. Category boundary and uncertainty margin are indicated (further, see text).

rises and falls in the time-course of the pitch of spoken sentences. Speech melodies can be parameterized cross-linguistically and described in much the same way as has been done for the segmentals in language: a set of distinctive features defines an inventory of abstract units, which can be organized in higher-order units subject to well-formedness constraints. Moreover, intonational contrasts are used to perform functions that can also be expressed by lexico-syntactic means, such as turning statements into questions, and putting constituents in focus. For these reasons it has become widely accepted that intonation is part of the linguistic system (Ladd, 1996: 8). Yet, there have always been adherents of the view that speech melody should be considered as something outside the realm of linguistics proper, i.e., that intonation is a paralinguistic phenomenon at best, to be treated on a par with the expression of attitudes or emotions. Typically, the communication of emotions (such as anger, fear, joy, surprise) or of attitudes (such as sarcasm) is non-categorical: the speaker shows himself more or less angry, fearful, or sarcastic in a continuous, gradient fashion.

A relatively recent insight, therefore, is that a division should be made in melodic phenomena occurring in speech between linguistic versus paralinguistic contrasts. Obviously, only the former but not the latter type of phenomena should be described by the grammar and explained by linguistic theory. This, however, begs the question how the difference can be made between linguistic and paralinguistic phenomena within the realm of speech melody. Ladd & Morton (1997) were the first to suggest that the traditional diagnostic for categorical perception should also be applicable to intonational categories. Only if a peak in the discrimination function is found for adjacent members on a tone continuum straddling a boundary between tonal categories, are they part of the linguistic system, i.e., phonological categories. If no categorical perception of the tone categories can be established, the categories are 'just' the extremes of a paralinguistic or phonetic tonal continuum. Ladd & Morton tested the traditional diagnostic on a tone continuum between normal and emphatic accent in English and noted that it failed.

Remijsen & van Heuven (2003) tested the traditional diagnostic on a tone continuum between 'L%' and 'H%' in Dutch, and showed that indeed there was a discrimination peak for adjacent members along the continuum straddling the boundary. At the same time, however, we had to take recourse to listener-individual normalization of the category boundary, a complication that is not generally needed when dealing with contrasts in the segmental phonology. Moreover, our relatively weak categorical effects could have been the result of an incorrect subdivision of the 'L%' to 'H%' tone range. Van Heuven & Kirsner (2002) showed that Dutch listeners were perfectly able to categorize final pitches in terms of three categories, functionally denoted as 'command' intonation, 'continuation', and 'question' (see §3). However, we did not run the full diagnostic involving both identification and discrimination. Also, we asked our listeners to choose between three response alternatives, viz. command, conditional and question. Although the extremes of the range, i.e. command versus question, are unchallenged linguistic categories, the conditional may well be non-distinct from the question type. After all, in the grammar developed by 't Hart, Collier & Cohen (1990) any type of non-low terminal pitch falls into the same category, indicating non-finality. It occurred to us that we should take the precaution to run the experiment several times, using different response alternatives, such that two separate binary ('command' ~ 'no command' and 'question ~ 'no question') response sets as well as our original ternary response set ('command' ~ 'conditional' ~ 'question') were used by the same set of listeners. If the intermediate 'conditional' response category does constitute a clearly defined notion in the listeners' minds, the binary and ternary divisions of the stimulus range should converge on the category boundaries. The present paper seeks to remedy our earlier infelicities.

We conclude this introduction by summarizing our research questions:

1. Are there three phrase-final boundary tones in Dutch: low for 'command' (also used for 'statement'), intermediate for 'conditional', and high for 'question' or just two, lumping the latter two together into a single category 'non-terminal'?
2. Where along the continuum are the category boundaries between the two or three boundary tones?
3. Are the category boundaries at the same positions irrespective of the binary versus ternary response mode?
4. Are the boundaries truly categorical in the sense that there are discrimination peaks for adjacent stimulus pairs straddling the category boundaries?

2. Methods

A male native speaker of standard Dutch read the sentence *Neemt u de trein naar WAgeningen?* with a single 'H*L' accent on the first syllable of *Wageningen.* The utterance was recorded onto digital audio tape (DAT) using a Sennheiser MKH 416 unidirectional condenser microphone, transferred to computer disk (16 kHz, 16 bits) and digitally processed using the Praat speech processing software (Boersma & Weenink, 1996). The intonation pattern of the utterance was stylized by hand as a sequence of straight lines in the ERB by linear time representation. Nine intonationally different versions were then generated using the PSOLA analysis-resynthesis technique (e.g. Moulines & Verhelst, 1995) implemented in the Praat software. The nine versions were identical up to and including the 'H*L' configuration on *Wageningen.* From that point onwards the nine versions diverged into two falls and seven rises. The terminal frequencies of the nine versions were chosen to be perceptually equidistant, i.e., the difference between any two adjacent terminal frequencies was equal in terms of the ERB scale.[1] The terminal pitch of version 1 equaled 80 Hz, the increment in the terminal frequency for each following version was 0,25 ERB, as shown in Figure 2.

For the **discrimination** task, which was the first task imposed on the subjects, stimuli were presented in pairs that were either the same or one step apart on the continuum. In the latter case, the second can be higher or lower than the first (hereafter AB and BA, respectively). The eight AB stimulus types ran from pair {1,2} to {8,9}; the eight corresponding BA types from {2,1} to {9,8}. This yielded 9 identical pairs and $2 \times 8 = 16$ different pairs, which occurred in random order, yielding a set of 25 trials in all, which was presented to each listener four times in different random orders, preceded by five practice trials. Stimuli within pairs were separated by a 500-ms silence, the pause between pairs was 3000 ms. For the **identification** task listeners responded to individual stimuli from the 9-step continuum by classifying each either in terms of a binary or a ternary choice:

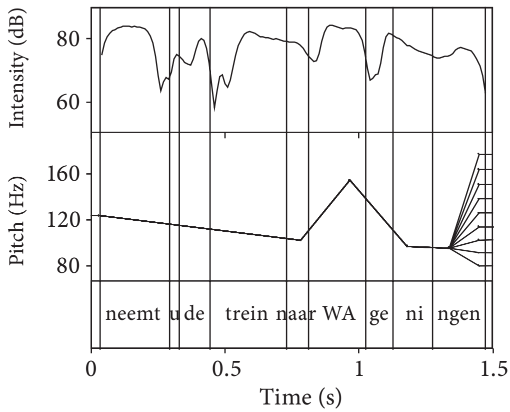

Figure 2. Steps 1 through 9 along resynthesized continuum differing in terminal pitch by 0,25 ERB increments.

(1) 'Command' ~ 'no command'. In one task the listeners were instructed to decide for each stimulus whether they interpreted it as a command or not.
(2) 'Question' ~ 'no question'. An alternative task involved the decision whether the stimulus sounded like a question or not.
(3) 'Command' ~ 'condition' ~ 'question'. The third task was identical to the task imposed in van Heuven & Kirsner (2002).

Half of the listeners first performed task (1), the other half of the listeners began with task (2). Task (3) was always the last identification procedure in the array of tests. For each task, the set of nine stimuli were presented five times to each listener, in different random orders, and preceded by five practice items, yielding sets of 50 identification stimuli per task.

Ten male and ten female Dutch listeners took part in the experiment on a voluntary basis. Participants were university students or members of their families. Subjects listened to the stimuli at a comfortable loudness level over Quad ESL-63 electrostatic loudspeakers, while seated in a sound-treated lecture room. They marked their responses on printed answer sheets, always taking the discrimination task first and the identification tasks last.

3. Results

Figures 3 and 4 present the results obtained in the binary identification tasks, i.e., the forced choice between 'command' ~ 'no command' (Figure 3) and between 'question' ~ 'no question' (Figure 4). The psychometric function for the 'command' responses is very steep. The category boundary between 'command' and 'no command' is located at a step size of 2.7, and the margin of uncertainty runs between 2.2 and 3.7, i.e., a cross-over from 75% to 25% 'command' responses is effected by an increase in the terminal pitch of the stimulus of 1.5 step (i.e., 0.37 ERB). A complete cross-over is also found for the 'question' ~ 'no question' task. The category boundary finds itself at a stimulus value of 3.6, whilst the margin of uncertainty runs between 2.3 and 4.9, i.e., an interval of 2.6 increments of 0.25 ERB. The category boundaries in the 'command' and 'question' tasks do not coincide, but diverge by almost a complete step: 2.7 versus 3.6 step. Note, once more, that none of the subjects had been alerted to the possible existence of an intermediate category between 'command' and 'question'. Therefore, the emergence of the interval between the 'command' and the 'question' boundaries might be taken in justification of an intermediate category.

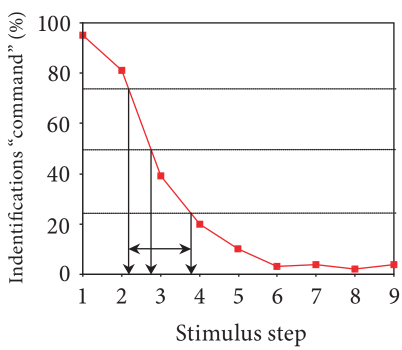

Figure 3. Percent 'command' responses as a function of stimulus step in binary identification of 'command'~'no command'.

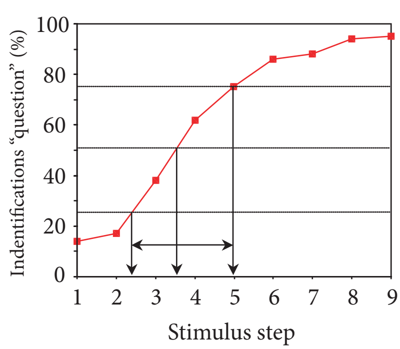

Figure 4. As Figure 3, but for 'question' ~ 'no question'

Let us now turn Figure 5 for the results of the ternary identification task, in which the same listeners were required to classify the nine stimulus types as either 'command', 'conditional subclause' or 'question'. The boundary between 'command' and 'continuation' is at 2.8; this is hardly different than the 'command' ~ 'no command' boundary that was found in the binary response task. This, then, would seem to be a very robust boundary, showing that at least 'command' intonation has well-defined linguistic status. The boundary between 'continuation' and 'question' is less clearly defined. Also, the maximum scores in these two categories are around 80% rather than 90% or more. Although there is no ambiguity in the listeners' minds whether a stimulus is a command or something else, the choice between 'continuation' and 'question' seems more ambiguous, leaving room for a minority response in the order of 20%. This would indicate to us that we are dealing here with a continuum rather than with a dichotomy. Finally, we may note that the (soft) category boundary between 'continuation' and 'question' is located at a stimulus value of 7.2. The boundary, then, that sets off 'question' from 'no question' responses proves unstable: there is a shift from the binary response task (3.6) to the ternary task (7.2) of no less than 3.6 points along the stimulus continuum. It would seem, then, that the 'command' category is highly stable and well-established in the minds of the listeners. The 'question' boundary, however, is rather poorly defined.

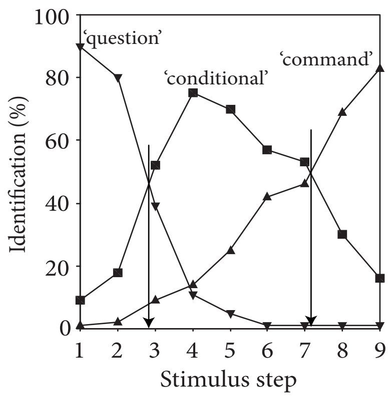

Figure 5. Ternary identification of stimuli as 'command', 'conditional clause' or 'question'. Category boundaries are indicated.

Figure 6 presents the mean percentage of successfully discriminated stimuli that were actually different (hereafter 'hits'), and the percentage of false alarms, i.e. 'different' responses to (identical) AA stimuli. The false-alarm rate is roughly 20% across the entire stimulus continuum. This value can be seen as a bias for responding 'different'. Generally, an increment of 0.25 ERB is discriminated above the 20% bias level, with the exception of the difference between stimulus steps 5 and 6. The discrimination function shows two local peaks. The first one is very large, and is located between stimulus steps 2 and 3. This peak obviously coincides with the stable category boundary found between 'command' and the non-command responses (whether binary or ternary). A

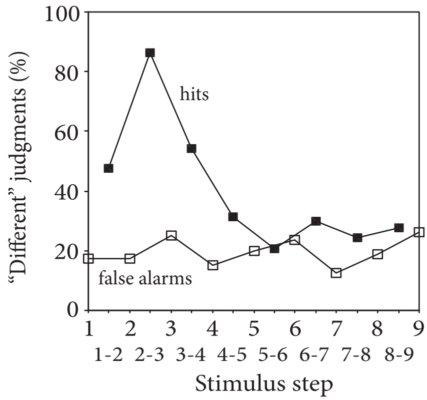

Figure 6. Percent 'different' judgments to nine identical stimulus pairs (false alarms) and eight pairs differing by one step (hits).

much smaller second discrimination peak may be observed between stimulus steps 6 and 7, which location may well reflect the rather poorly defined category boundary between 'continuation' and 'question'.

Using the so-called Haskins formula $D_{pred} = ½ \times (1 + (P_1 - P_2)^2)$ to predict correct discrimination of different stimulus pairs from the identification scores obtained for each of the stimulus steps along the continuum, we find very high correlations between predicted and observed discrimination for 'command' identification: $r = .949$ ($p < .001$) in the ternary and $r = .932$ ($p = .001$) in the binary task. Prediction of discrimination behavior from 'question' identification is only just significant in the binary task: $r = .710$ ($p = .049$) but negative (and insignificant) in the ternary identification task: $r = -.545$ ($p = .163$). These statistics, again, suggest that only the 'command' ~ 'no command' contrast is categorical.

4. Conclusions and discussion

Let us now try to formulate answers to the research questions in §1.2. The first question was whether the domain-final boundary tones are contiguous categories along a single tonal dimension, and map onto the command, continuation and question meaning in a one-to-one fashion. The results of our experiments clearly indicate that this is indeed the case. Our listeners had no difficulty in using the three response alternatives provided to them. When the terminal pitch was lower than the preceding pivot point in the contour the responses were almost unanimously for 'command'. When the final pitch was higher than the preceding pivot point, the incidence of 'continuation' responses increased up to and including step 4, and decreased for higher terminal pitches which were more readily identified as questions as the terminal pitch was higher.

Although there was always some ambiguity between the 'continuation' and 'question' alternatives, the results clearly indicate that 'continuation' is signaled by moderate final pitch, and question by (extra) high pitch.

The latter finding suggests that asking a question involves a higher degree of appeal by the speaker to hearer than asking the listener's continued attention. We may also note that our result clashes with Caspers (1998). She found that the intermediate final pitch (or high level pitch in her experiment) was unambiguously identified as continuation; extra high final pitch ambiguously coded either continuation or question. Comparison of Caspers' and our own results is hazardous since the utterance-final tone configurations differ, not so much at the underlying tone level, but at the surface. It seems to us that the discrepancy between Caspers' and our own findings can be resolved if we accept the possibility that Caspers' extra high terminal pitch was simply not high enough to elicit the 80% 'question' responses that we got in our experiment.

The results so far concur with van Heuven & Kirsner (2002). However, we may now go on to consider the second, third and fourth questions, which asked where the category boundaries are located along the final pitch continuum between low, intermediate and high, in the binary and ternary response tasks, and to what extent the boundaries coincide with a peak in the discrimination function. The results obtained in the binary ('command' ~ 'no command') and ternary ('command' ~ 'continuation' ~ 'question') identification tasks are virtually the same, yielding the same location of the boundary (at step 2.7) separating the 'command' category from the rest of the stimulus continuum. However, a very unstable boundary is found in the binary 'question' ~ 'no question' task (at step 3.6), which is reflected in the poorly defined boundary separating the 'continuation' and 'question' categories in the ternary response task (at step 7.2). Moreover, we have seen that the category boundary between 'command' and 'no command' coincides with a huge peak in the discrimination function. Although there is a modest local maximum in the discrimination function that may be associated with a boundary between 'continuation' and 'question', this peak is not very convincing.

We take these findings as evidence that there is a linguistic, or phonological, categorization of the final boundary tone continuum in just two types, which is best characterized as low and non-low. The low boundary tone signals dominance or superiority on the part of the speaker. This is the boundary tone that is suited for issuing statements and commands. The non-low boundary tone signals subservience of the speaker to the hearer; the speaker appeals to the hearer for his continued attention or for an answer to a question.

The non-low part of the boundary opposition, however, represents a gradient, paralinguistic continuum between a moderate appeal (asking for the hearer's continued attention) and a stronger appeal (asking the hearer for a verbal reply to a question). Here the lower terminal pitches are associated with weaker degrees of appeal (or subservience), and the higher levels with strong appeal, but in a continuous, gradient, non-phonological manner.

Our results indicate that earlier findings reported by Remijsen & van Heuven (2003) are to be viewed with caution. We now know that the listeners' task should not be to decide whether the stimulus is a statement (or a command) versus a question. If binary response alternatives are required, the categories should be 'statement' versus 'no statement' but a better procedure would be to ask the listener to choose from three categories: 'statement' (equivalent to 'command' in our experiments) ~ 'continuation' ~ 'question'. Had such precautions been taken by Remijsen & van Heuven, their category boundary would have been much better defined with less listener-individual variation.

Methodologically, the classical identification-cum-discrimination paradigm is a useful diagnostic tool in intonation research which allows linguists to decide experimentally whether a melodic contrast is categorical and therefore part of the phonology, or continuously gradient and therefore phonetic or even paralinguistic.

Notes

1. The experiments reported in this article were run by Suzanne Strik and Josien Klink in partial fulfillment of the course requirements for the Experimental Phonetics Seminar 2003/04 taught by the Linguistics Programme at Universiteit Leiden.

2. The ERB scale (Equivalent Rectangular Bandwidth) is currently held to be the most satisfactory psychophysical conversion for pitch intervals in human speech (Hermes & van Gestel, 1991). The conversion from hertz (f) to ERB (E) is achieved by a simple formula: $E = 16.6 \times \log(1 + f/165.4)$.

References

Boersma, P. and Weenink, D. (1996). Praat, a system for doing phonetics by computer. *Report of the Institute of Phonetic Sciences Amsterdam*, 132.

Caspers, J. (1998). 'Who's next?' The melodic marking of question vs. continuation in Dutch. *Language and Speech* 41, 375–398.

Hart, J. 't, Collier, R. and Cohen, A. (1990). *A Perceptual study of intonation. An experimental-phonetic approach to speech perception.* Cambridge: Cambridge University Press.

Hermes, D.J. and Gestel, J.C. van (1991). The frequency scale of speech intonation. *Journal of the Acoustical Society of America* 90, 97–102.

Heuven, V.J. van and Kirsner, R.S. (2002). Interaction of tone and particle in the signaling of clause type in Dutch, in H. Broekhuis, P. Fikkert (eds.). *Linguistics in the Netherlands 2002*, Amsterdam/Philadelphia: John Benjamins, 73–84.

Ladd, D.R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Ladd, D.R. and Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics*, 25, 313–342.

Moulines, E. and Verhelst, E. (1995). 'Time-domain and frequency-domain techniques for prosodic modification of speech'. In W.B. Kleijn and K.K. Paliwal, eds., *Speech coding and synthesis.* Amsterdam: Elsevier Science, 519–555.

Remijsen, A.C. and Heuven, V.J. van (2003). Linguistic versus paralinguistic status of prosodic contrasts, the case of high and low pitch in Dutch, in J.M. van de Weijer, V.J. van Heuven, H.G. van der Hulst (eds.): *The phonological spectrum. Volume II: Suprasegmental structure.* Amsterdam/Philadelphia: John Benjamins, 225–246.