Toni Rietveld & Aoju Chen (Nijmegen)*

How to Obtain and Process Perceptual Judgements of Intonational Meaning

1 Introduction

1.1 Intonational meaning

This chapter focuses on research into intonational meaning in Germanic languages. An overview of the current understanding of intonational meaning in these languages is thus in order (see Chen, 2005 for a detailed literature review). Generally, there are two aspects of intonation that convey meaning, the pitch contour and its phonetic implementation.[1] Analyses of intonational meaning signalled by the pitch contour can vary in their emphasis as to which is semantically more relevant, the whole contour or its parts (Ladd, 1978). In tunes-based analyses, the meaning of the pitch contour is believed to be mainly conveyed by the whole contour. For example, four whole-utterance contours arise from the Liberman-Sag model (Liberman and Sag, 1974; Sag and Liberman, 1975), i.e. the 'contradiction-contour', the 'tilde-contour', which makes a wh-question an unambiguously real question (as opposed to a rhetorical one), the 'hat-contour', whereby a wh-question can be interpreted both as a real question and as a negative-implicating rhetorical question or suggestion, and the 'surprise/redundancy contour'. In tones-based analyses, the meaning of a pitch contour is assumed to stem mostly from the meanings of its parts. For example, Pierrehumbert and Hirschberg (1990) propose that the meaning of the contour is compositionally derivable from its components, i.e. pitch accents (H*, L*, H*+L, H+L*, L*+H, L+H*), phrase accents (H-, L-), and boundary tones (H%, L%). As regards phonetic implementation, a pitch contour can be realised with different peak and valley alignments and pitch ranges. Alignment refers to the relative timing of the pitch peak or valley in segments. Pitch range

* We would like to thank Ingrid Schiks for her help with recruiting participants, Femke Deckers and Femke Uijtdewilligen for administering Experiment 2 and preparing the data for statistical analyses, Carmel O'Shannessy for her advice on wording and stylistic issues, and our participants for their patience and cooperation.

1 The term 'pitch contour' is used here to refer to the phonological pitch contour, which can have numerous phonetic variants resulting from phonetic implementation. Related to pitch contour, prosodic phrasing (i.e. the division of utterances into smaller phrases) also signals meanings, usually together with the placement of pitch accent.

can be varied along two dimensions, span and register (Ladd, 1996, pp. 260-261, Cruttenden, 1997, pp. 123-124). Span variation involves increases or decreases in the distance between the highest and the lowest pitches in the contour; register variation involves raising or lowering of both the high and the low pitches (Gussenhoven, 1999). Pitch range variation can also be local, e.g. changes in end pitch height.

The meanings that pitch contours have been shown to convey are as follows: grammatical meaning (e.g. sentence type) (Halliday, 1967); pragmatic usage of sentence type (Sag and Liberman, 1975); attitude of the speaker (e.g. Pike, 1945; O'Connor and Arnold, 1973), and discoursal meaning. The notion of discourse meaning has been operationalised as the relation between the variable in focus and an implicit set of variables in the context (Ladd, 1978), the status of the information carried in a particular utterance with respect to a 'background' created in the course of the exchanges between the speaker and the hearer or hypothesised by the speaker (Gussenhoven, 1984), or the relationship between the propositional content of a particular intonational phrase and the mutual beliefs of the speaker and the hearer (Pierrehumbert and Hirschberg, 1990). It is now generally accepted that the relation between intonation and grammar is 'casual not causal' in that 'grammar uses intonation on those frequent encounters, but intonation is not grammatical' (Bolinger, 1965, p. 100). Attitudinal approaches are problematic because the division between attitudinal and grammatical meaning is lacking in practice. Tune meaning can thus be more satisfactorily analysed at the level of discourse.

Pitch range variation is traditionally believed to signal speaker attitude and emotion. In a recent analysis, Gussenhoven (2002) distinguishes two types of meaning. One is informational, concerned with attributes of the message. For example, final high pitch signals continuity and final low pitch signals finality. The second is affective, concerned with attributes of the speaker. For example, a high register conveys submissiveness, whereas a low register conveys dominance.

Having considered the forms that can carry intonational meaning and the types of meaning that intonation can signal, we now briefly discuss how form can be associated with meaning. A distinction may be made between discrete and gradient form-function relations. For the discrete form-function relation, a change in form leads to a categorical change in meaning, e.g. H*L conveys that the associated lexical item conveys information new to the discourse while L*H conveys that the associated lexical item conveys information already present in the discourse. For the gradient form-function relation, a change in form leads to a change in the degree of a certain meaning, e.g. a wider span signals a higher degree of emphasis. The former is known as the linguistic signalling of intonation, and the latter the paralinguistic signalling of intonation. The pitch contour and its parts are assumed to be relevant for linguistic intonational

meaning, while the phonetic implementation is assumed to be relevant for paralinguistic intonational meaning.

As argued by Gussenhoven (2002), linguistic form-function relations are language-specific and can be arbitrary; paralinguistic form-function relations are derived from physiological conditions that are responsible for within- and between-speaker pitch variations and are thus universal. However, cross-linguistic perception research (Chen, 2005) showed that speakers with different language backgrounds differ in their perception of paralinguistic form-function relations as a result of various factors, including standard pitch range and linguistic intonational meaning in their native languages. For instance, speakers of a language with a narrower standard pitch range (e.g. Dutch) appear to perceive greater meaning differences across a given pitch span than speakers of a language with a wider standard pitch range (e.g. British English).

It should be noted that the line between linguistic and paralinguistic intonational meaning is not always clear-cut for two reasons. First, gradient form-meaning relations can be grammaticalised in such a way that they become discrete. For instance, a higher end pitch sounds more questioning, which is grammaticalised in many languages such that H% is associated with questions but L% with statements. Second, variations in span and peak alignment can lead to a discrete change in meaning. A case in point is provided by Kohler (1987), who showed that, in German, the fall contour with an early peak is perceived to signal the meaning 'established' while the fall contour with a late peak is perceived to signal the meaning 'new'. Similarly, in the Neapolitan variety of Italian, early peak alignment signals statements while late peak alignment signals questions. The difference between early and late peak alignments is about 40ms (e.g. D'Imperio and House, 1997). As for pitch span, Ladd and Morton (1997) showed that a very high peak is perceived to signal emphasis but a normal high peak is not associated with this interpretation in English.

In empirical studies of intonational meaning, the use of a certain perceptual scale implicitly suggests that the perception is measured in a gradient way. Nevertheless the data can be used as evidence in favour of or against a postulated discrete form-function relation (see Section 2 of this chapter for more discussion). The issue of whether a certain pair of intonational contrasts, which may or may not be accompanied by categorical differences in meaning, are discrete or gradient is a separate research question. A number of methods have been put forward to determine the nature of intonational contrasts. See Gussenhoven in this volume for a review of available methods.

1.2 Why perceptual studies on intonational meaning?

Focusing on perception in studies of intonational meaning can be either methodologically motivated or driven by a research question. As a method, via the

use of manipulated speech, perceptual studies allow controlled modification of intonational parameters that may contribute to a certain meaning, and are hence suitable for establishing form-meaning relations. In combination with a descriptive approach to intonational meaning, they can help shed light on how the cues available in the segments contribute to the conveyance of the meaning. As a research question, investigators have been concerned with how a certain intonational parameter is interpreted in different pragmatic or discoursal contexts by listeners with the same or different language backgrounds.

2 Perceptual scales

A large number of methods are available to obtain perceptual judgements of attributes of intonation contours. Perceptually judging an object is a way of measuring the object. Measuring is assigning numbers or categories to objects in such a way that relations between the objects are reflected in the relations between the numbers or categories. The kind of relations that can be reflected in the numbers or categories depends on the properties of the scale on which the objects are projected. The word 'scale' can be interpreted in different ways. It can refer to a line or a roll of boxes available for listeners to mark a value, corresponding with the extent to which an object has a specific property, as in the Equal Appearing Interval Scale and the Visual Analogue Scale. It can also refer to the result of statistical manipulations of raw data that are supposed to yield positions of experimental conditions on a line. These positions reflect the mutual distances between experimental conditions in an optimal way. For example, having obtained data by means of the Paired Comparisons method on the perception of friendliness as signaled by pitch register, statistical manipulations can yield positions of the three register conditions (i.e. A, B, and C) on a scale, as illustrated in Figure 1.

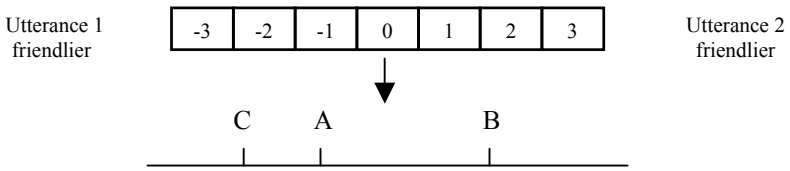

Figure 1: Projection of values obtained via paired comparisons for experimental conditions A, B, and C on a final scale after statistical manipulation

The scales mentioned above are *unidimensional*, that is, they deal with only one attribute of the objects at issue, for instance the degree of friendliness. Section 3.1 is devoted to the use of these scales. There are also tasks that involve

more global judgements, for example, 'the dissimilarity between the intonation of utterance A and the intonation of utterance B is 3 points' (on a scale of 10 points), or 'the intonation of utterances A and B differs less than the intonation of utterances C and B'. In most of these cases, a unidimensional scale does not suffice to represent the dissimilarities δ_{ij} between objects i and j, as (dis)similarity judgements are often based on more than one attribute (see Section 3.2 for further discussion).

In view of the second usage of the word 'scale', there are four scale types, i.e. the nominal scale, the ordinal scale, the interval scale, and the ratio scale. These scales can be characterised by the kind of relations between objects A and B that can be reflected and the kind of operations that are allowed on scale values.

1. *Nominal scale*. Relevant relations: A = B, or A ≠ B. Every one-to-one substitution of scale values is allowed, provided that the categorization of all objects as the 'same' or 'different' is not affected. This means that one may categorize utterances as 'friendly' or 'unfriendly', but the transformation into 'A' or 'B' is allowed. The results from statistical procedures carried out on the data (often counts or frequencies) will not be affected.

2. *Ordinal scale*. Relevant relations: A = B, A ≠ B, A > B. Every monotone transformation is allowed. A monotone transformation is one that leaves the ordering of objects unaffected. Thus a transformation like $f(x) = a \times x^2$ is allowed. For example, if the scale values 1, 3 and 10 are transformed according to $a \times x^2$ we get (with $\alpha = 2$) 2, 18, and 200, respectively. The ordering of the objects remains the same. It is not easy to find examples in phonetics that are clearly measured on an ordinal scale. A number of subjective judgements are said to be measured on an ordinal scale, as we cannot be completely certain that differences between judgements with the values 3 and 4 on one hand, and between judgements with the values 6 and 7 on the other hand are equal, which would be equal in a measurement on the interval scale.

3. *Interval scale*. Relevant relations: A = B, A ≠ B, A > B, (A – B) = (C – B), or (A – B) > (C – B). The linear transformation $f(x) = a \times x + b$ is allowed. Scores on an interval scale may be multiplied by a constant, and a constant may be added, as the zero point has no meaning. Intervals of magnitude *m* are always the same along the whole scale, e.g. when comparing intrinsic vowel durations, the difference between 60ms and 80ms is the same as the difference between 200ms and 220ms. Notice that the intervals are the same in a physical sense, but not necessarily in a perceptual sense. Take for example the perceived pitch peak measured in Hz. The difference between 100Hz and 110Hz is perceptually larger than the difference between 400Hz and 410Hz. The absolute duration of a sound is clearly not measured on an ordinal or interval scale, as zero has a clear meaning in that the associated sound is not realised.

4. *Ratio scale*. Relevant relations: A = B, A ≠ B, A > B, (A – B) = (C – B), (A – B) > (C – B), A/B = C/B. Only the transformation $f(x) = a \times x$ is allowed;

adding a constant would make the zero point disappear. Examples of objects which are measured on a ratio scale are sound pressure levels, scaled either as Pascal and N/m^2. A value of 0 has a clear meaning: silence.

Statistical procedures are applied according to the scale type. For example, if the property of equality of intervals along the whole scale is not warranted, researchers tend to be reluctant to use statistical procedures that are based on these properties, such as a *t*-test, or an analysis of variance (hereafter ANOVA), which are known as the parametric statistics. This topic will be discussed in Section 4 of this chapter.

3 Scaling procedures and examples from intonational research

3.1 Unidimensional scaling

The following four unidimensional scaling procedures will be considered in four subsections.

1. Equal Appearing Interval scale (EAI)
2. Paired Comparisons (PC)
3. Direct Magnitude Estimation Measurement (DME)
4. Visual Analogue Scale (VAS)

In each subsection, we discuss, after a short introduction, the characteristics of the scaling procedure, the corresponding method of data processing, the procedure's advantages and disadvantages, followed by recommendations on how to use it and examples of applications in intonational research.

3.1.1 Equal Appearing Interval scale (EAI)

The Equal Appearing Interval scale was developed by Thurstone (1928) for the measurement of attitudes towards disputed social issues. It is also known as the Thurstone scale. The procedure consists of generating a large number of potential scale items (i.e. statements about a particular issue), having judges rating each statement on a scale (e.g. 1 to 11) in terms of the extent to which each statement indicates a favourable attitude, computing scale score values (i.e. median and the interquartile range), selecting the final scale items (i.e. the statements that are at equal intervals across the range of medians), administering the scale (i.e. asking another group of judges whether they agree or disagree with each statement), and finally calculating the total scale score for each judge by averaging the scale scores of all the items that the judge agrees with.

An adjusted version of this method has been used in many other contexts, as shown in Figure 2. Here, the researcher presents participants with a number of utterances (which were composed to represent various experimental condi-

tions) and asks them to judge for each utterance how friendly the speaker sounds, on a scale from 1 to 7 by ticking the appropriate box. '7' stands for 'very friendly'; '1' stands for 'not friendly'. The assumption is that the difference in degree of friendliness between utterances with the score 2 and utterances with the score 1 is the same as the difference between utterances with the score 7 and utterances with the score 6. In the literature, the adjusted version of the EAI scale is conventionally referred to as the EAI scale. In what follows, we will adhere to this convention. Note that researchers often differ in the number of categories they use, with 7 being the most frequently used number. In fact, there is little literature on the number of categories to be used in the EAI scale. Exceptional is Jensen and Tøndering (2005). They compared EAI scales with 2, 4 and 31 categories in the perception of prosodic prominence. It was found that the two-point and four-point EAI scales clearly outperformed the 31-point scale with the four-point scale performing slightly better than the two-point scale.

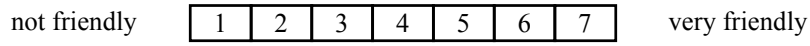

Figure 2: An example of the Equal Appearing Interval Scale

A scale that is sometimes confused with the EAI scale in the literature is the Likert Scale (Likert, 1932). Differing from the original EAI scale, which asks the judges to indicate their (dis)agreement with multiple statements over a single issue, the Likert Scale asks the judges to indicate their (dis)agreement with a single statement over a single issue typically on a scale ranging from 1 to 5. Often the scale will be 1 = strongly disagree, 2 = disagree, 3 = not sure, 4 = agree, and 5 = strongly agree, as illustrated in Figure 3.

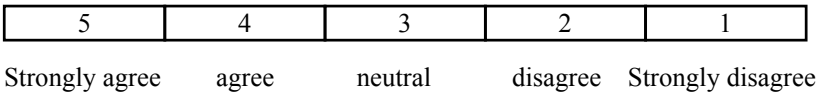

Figure 3: An example of the Likert Scale

The differences between the EAI scale and the Likert Scale are, however, subtle. Take for example the rating of friendliness. In the EAI scale, judges judge directly the degree of friendliness. In the Likert Scale, judges would indicate to what extent they agree with the statement 'The speaker sounds friendly'. Because a higher agreement score can be interpreted to mean a higher degree of friendliness, we may say that the Likert Scale measures the degree of friendliness in an indirect way.

Procedure:

Judges are asked to tick the box that corresponds most with the extent to which an object has a certain characteristic.

Processing:

In general, the judgements given by the judges are averaged; thus a panel judgement per object is obtained and analysed with parametric tests.[2]

Advantages:

1. It is easy to administer and process.
2. Judges get a clear idea of the task.

Disadvantages:

1. The prescribed nature of the EAI scale may not capture the judge's full range of perception (Stevens, 1974). If, for instance, there are eight gradations in degree of prominence, the restriction of a scale to seven categories makes the judge's task less easy.
2. In some sensory dimensions, judges appear to partition the lower end of the continuum into smaller intervals than at other locations on the scale (Stevens, 1974)
3. Judges are also reported to have the tendency of assigning stimuli to categories in such a way that all scores are used equally often (Gescheider, 1976).
4. Inexperienced judges, however, tend to use the middle category more often than the extremes.

Recommendations:

1. Make sure that not too many scales are printed on one page or shown on the computer screen.
2. Present an anchor stimulus, which has approximately the middle value of the scale, at a regular interval. This is done to maximise consistency within judges.
3. Do not change the order in which the two ends of the scale are presented, e.g. *bad – good*, *less – more*, in a test, as it might confuse judges.

[2] Usually, the analyses are preformed on the raw scores. When there is evidence suggesting an observable difference in the scoring styles between judges, it may be desirable to first convert the raw scores to z-scores and then perform the analyses (see also Grabe, Gussenhoven, Haan, Marsi, and Post, 1997).

Examples of applications:

The EAI scale has been used with modifications both at the stage of obtaining judgements and at the stage of data processing. In the former case, judges do not make judgements by assigning a score but by selecting a response from a set of descriptive responses that reflect a continuum of a certain meaning; the descriptive responses are then converted into scores for the purpose of analysis. In the latter case, the EAI scores are only used to construct a new set of data, on which a statistical analysis is performed. In the next two paragraphs, we give one example for each case.

In a study on the effect of final fall (i.e. the Intonational-Phrase final pitch accent) on the perception of finality, Wichmann (1991) asked native speakers of British English to listen to syntactically complete sentences varied in the starting point of the final fall and to judge whether the speaker of each sentence was 'definitely going on', 'probably going on', 'had probably finished', or 'had definitely finished'. In the data processing stage, the response 'definitely going on' was assigned the score '1', 'probably going on' the score '2', 'had probably finished' the score '3', and 'had definitely finished' the score '4'. This procedure made it possible to calculate the 'finality' scores. It was found that the lower the starting point of the final fall, the higher the finality score.[3]

In a study on the discourse meaning of pitch accent type, Caspers (2000) verified meaning hypotheses for four Dutch pitch accents derived from Keijsper (1984) and Gussenhoven (1984). The four pitch accent types ('t Hart, Collier, and Cohen, 1990), including '1 2' (an accent-lending rise followed by a boundary-marking rise), 'A' (an accent-lending fall), '1&A' (an accent-lending rise and fall), and '1&E' (an accent-lending rise and a half fall), were realised on proper names, as illustrated in Figure 4. The meaning hypotheses are given in (1).

(1) The meaning hypotheses of Caspers (2000):

1 2	=	Testing:	The speaker leaves it up to the hearer to decide whether a variable belongs to the background;
A	=	Selection:	The speaker selects a variable from the background;
1&A	=	Addition:	The speaker adds a variable to the background;
1&E	=	Addition plus:	A variable is added to the background, but it is a matter of everyday routine.

[3] When evaluating the effect of the final fall statistically, the author treated the data as being ordinal and conducted a non-parametric test.

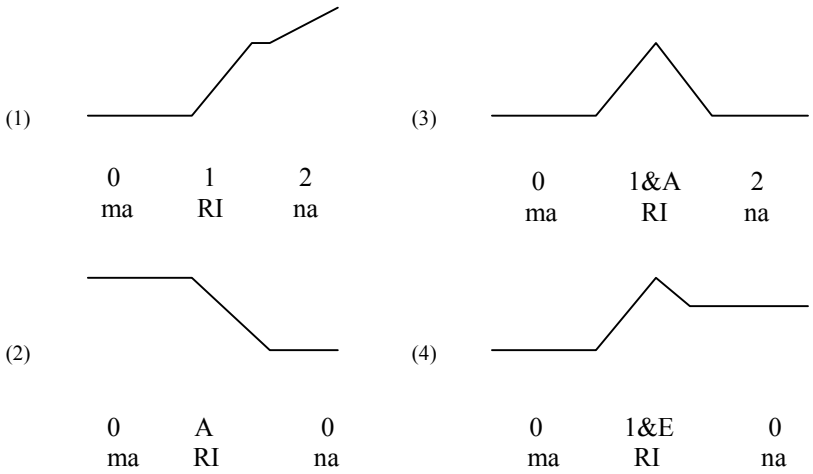

Figure 4: Stylised examples of the four pitch accents on the proper name 'Marina'. After Figure 1 of Caspers (2000, p. 131).

The hypothesised meanings were incorporated in two situational contexts (also referred to as orientations). In one orientation, the proper name was used to address a person (vocative context); in the other orientation, the proper name was simply the focused information (default context). An example of the vocative contexts is given in (2) with the accompanying proper name.

(2) Context utterance: You want to speak to a colleague about something important; this colleague, however, is in the staff room, talking to others. You join them and try:
Target utterance: *Marina*.

Native speakers of Dutch ranked the appropriateness of each pitch accent in each context on a four-point EAI scale ('1' for best fit to '4' for worst fit). For data processing, Caspers constructed a new set of data consisting of the frequency of each accent type as the most appropriate intonation (i.e. how many times an accent type was assigned the score '1') in each context, and performed an ANOVA with two within-subject factors, Meaning and Context. Note that the use of the EAI scale in the stage of obtaining judgements allows the judges more 'rating space' than does a forced-choice task (e.g. 'the intonation is appropriate in the context' vs. 'the intonation is not appropriate in the context'.), and can thus more accurately reflect how reliably a certain pitch accent signals a certain meaning.

Note also that in the use of the EAI Scale, the two ends of the scale are sometimes labelled with the negation of the adjective used for the other end (e.g. 'polite' vs. 'not polite'). It has been suggested that this kind of labelling

may not be appropriate. According to Grabe, Gussenhoven, Haan, Marsi, and Post (1997), it is potentially confusing, as the negative terms sometimes appear on the left and sometimes on the right, an arrangement that is generally recommended to ensure maximum independence of scales that have similar effects. Furthermore, the interpretation of the negative end is ambiguous. It can either be interpreted as expressing the presence of a nonfavourable attribute (e.g. 'the speaker is definitely impolite') or as expressing the absence of a favourable attribute (e.g. 'the speaker is not polite but is not necessarily impolite'). To circumvent these problems, Grabe et al. (1997) adopted the Likert Scale, i.e. expressing the meanings in full sentence statements (e.g. 'the speaker sounds polite') and asking listeners to rate scales that were labelled 'I agree' (with the statement) (score '1') vs. 'I disagree' (with the statement) (score '5'). A lower score thus indicated a higher degree of politeness. Similarly, Rietveld, Haan, Heijmans, and Gussenhoven (2002) expressed the meanings (pragmatic scales in their terminology) (e.g. 'surprised') with full sentence statements (e.g. 'by her intonation, the speaker indicates that she is surprised about the content') and asked listeners to rate how appropriate the contour was for a given meaning on a ten-point scale. Score '1' stood for 'the intonation contour does not match the given pragmatic scale at all'; '10' meant 'the intonation contour perfectly matches the given pragmatic scale'.

3.1.2 Paired Comparisons (PC)

Scheffé (1952) outlined a procedure to obtain an interval scale on which objects are located, based on a paired comparisons procedure. Paired comparisons involve a rather simple choice: is stimulus A more ... than stimulus B? Scheffé's approach combines the advantages of a binary choice with the possibility to give an answer on a scale. The technique has been used in the study of intonation but also in evaluation studies of synthetic speech (Eggen, 1992).

Procedure:

1. All stimuli are presented in pairs in both orders AB and BA.
2. One group of judges rates the stimuli in the order AB; another group of judges rates the stimuli in the order BA.
3. A scale of 7 or more points is adopted (see Figure 5 for an example).
4. Judges are asked to indicate which of the two items (A or B) in each pair has a higher degree of a certain attribute and how much higher it is.
5. 0 means: both items are characterised by the attribute to the same extent; -3: the first item of a pair is much 'better', etc.; +3: the second item of a pair is much 'better', etc.

6. For m objects, $m \times (m - 1)$ presentations of pairs are needed (e.g. 90 pairs for 10 objects); as each participant judges only one order of the stimuli, each participant is presented with $(m \times (m - 1))/2$ pairs.

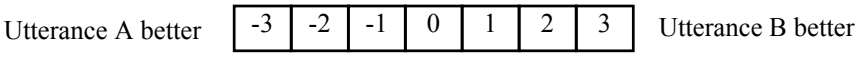

Figure 5: An example of a seven-point scale used in paired comparisons. Points -1, -2, -3 represent increasing degrees of preference for the first utterance. Points 1, 2, 3 represent increasing degrees of preference for the second utterance. Point 0 represents 'no preference'.

Advantages:

1. Binary judgements are easier to make than absolute judgements.
2. Gradations in judgements are possible.
3. The resulting scale is claimed to be at the interval level.
4. A yardstick is given, which enables the researcher to assess pairwise whether objects differ significantly.

Disadvantages:

1. Each participant has to judge a large number of stimuli; for example, a set of 15 stimuli results in 105 ($(m \times (m - 1))/2$) pairs of stimuli.
2. Order effects are possible ('2nd stimuli are ...er than 1st stimuli').
3. Dedicated software is needed to run the analysis, such as DIFANOVA (which is available from the first author, written in C, and run under MS-DOS and Command Prompt).
4. Repeated measures cannot be handled easily; the researcher needs to perform separate analyses for each implementation of an experimental condition.

Recommendations:

Randomise the orders of pairs for the judges. If there is a kind of 'natural' order of the stimuli A, B and C (e.g. level of pitch register), make sure that the first group has to judge the order AB, and the second group BA, but that the order AC is presented to the first group and CA to the second, and so on.

Examples of applications:

Chen (to appear) adopted Scheffé's paired-comparisons paradigm to establish which contour was perceived to be the preferred pitch contour to signal continuation at clause boundaries in British English, German and Dutch (Chen's Experiment 1). Intonationally different renditions of a compound sentence (e.g. *The story is too long but is fun to read.*) or a sentence sequence, (i.e.

a sequence of two simple sentences), (e.g. *The story is too long. The plot is boring.*) were presented in pairs to listeners in their native languages. Six contour-conditions (3 contours × 2 variants) (see Figure 6) were realised on four compound sentences and four sentence sequences.

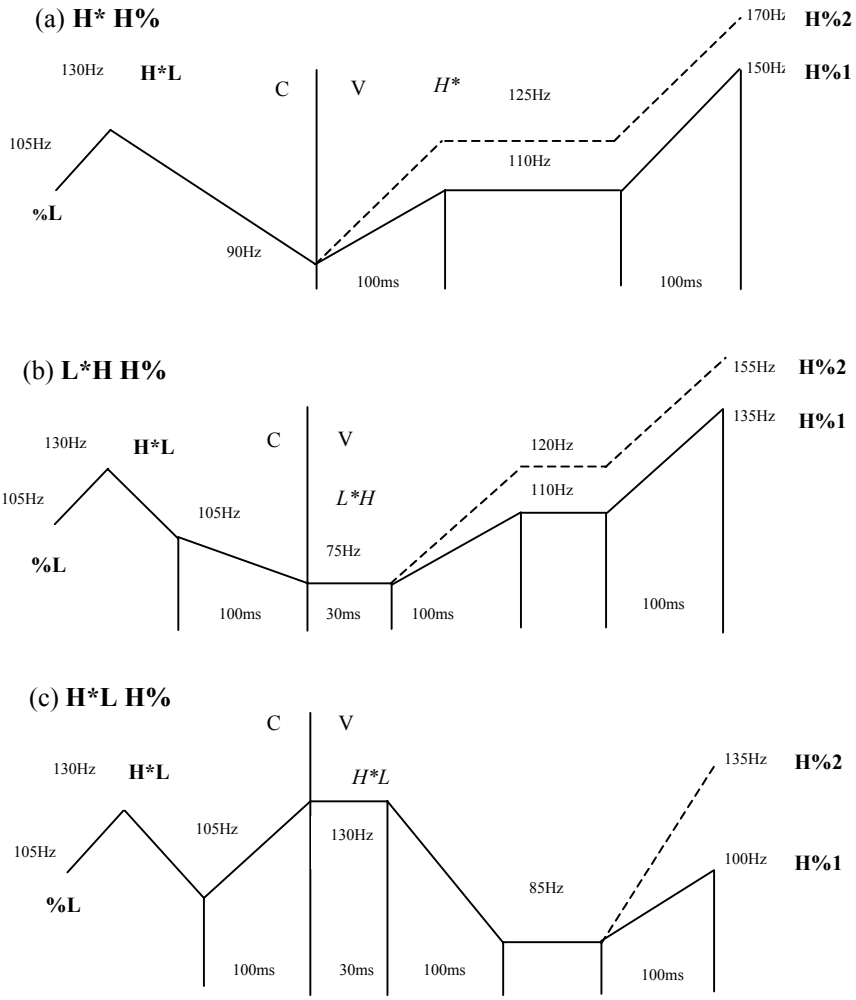

Figure 6: Schematic representations of %L H*L H* H% (abbreviated H* H%), %L H*L L*H H% (abbreviated L*H H%), and %L H*L H*L H% (abbreviated H*L H%) in non-IP final positions (i.e. the pitch accent is realised on a stressed syllable followed by at least another unstressed syllable) with variations in the final rising portion

There were thus six renditions per compound sentence/sentence sequence, resulting in 15 pairs in the order AB and another 15 pairs in the order BA. Listeners were asked to judge on each stimulus trial which of the two renditions sounded better in terms of how the two clauses in each rendition were intonationally connected, and indicate the degree to which this was the case on a seven-point scale (-3, -2, -1, 0, 1, 2, 3), as shown in Figure 5.

Note that Chen's design involved repeated measures. Since Scheffé's (1952) analysis cannot process repeated measures, data from each group of listeners were divided up into eight sets, with each set including only data obtained from the lexically identical stimulus pairs. Twenty-four separate analyses of variance for paired comparisons were then performed on the frequencies of scores per contour-condition pair, one for each data set of each listener group. Additional statistics (e.g. one-tailed Pearson's correlation coefficients) were obtained either to assess the generalisability of the findings emerging from the analyses of variance or to bring out subtler differences between contour-conditions.

3.1.3 Direct Magnitude Estimation measurement (DME)

Direct Magnitude Estimation measurement is a technique used to determine quickly and easily to what degree a person is experiencing a given sensation. It is a direct scaling method, by which judges are asked to directly judge the strength of a sensation induced by a stimulus (Allard, 2001).[4] Stevens and Galanter (1957) were the first experimenters to suggest using magnitude estimations to scale sensation quantitatively. In DME judges are asked to focus on ratios: "if the modulus is assigned the value of 100, how many times more or less is the magnitude of an attribute of the test stimulus than that of the modulus?" The ratios can result in, for instance, 300 ("3 times more") or 33 ("3 times less"). DME has been extensively used in psychophysics in order to establish the relationship between physical characteristics of stimuli (like the fundamental frequency of a sound, expressed in Hz) and their perceptual correlates ('pitch'). Psychophysicists distinguished two classes of dimensions: a) *metathetic* dimensions, and b) *prothetic* dimensions (Stevens, 1974). A *metathetic* dimension is a dimension which varies in terms of a 'change in quality'. Pitch is often considered a metathetic dimension. Variations along a *prothetic* dimension occur in 'degrees' of quantity or magnitude. Loudness is an example of prothetic dimensions. According to Stevens, a prothetic dimension is not suited for linear partitioning; as a consequence EAI is not to be recommended for scaling continua which should be regarded as prothetic dimensions. It follows that metathetic dimensions can be scaled by both the EAI

[4] See *http://ahsmail.uwaterloo.ca/kin356/magest/magest.htm* for an online tutorial on the Direct Magnitude Estimation measurement.

scale and the DME measurement. The relation between EAI and DME scale values provides a test to establish whether a dimension is to be seen as metathetic or prothetic. If these scale values – obtained for a specific attribute – are not linearly related, one has to assume a prothetic continuum for the attribute at issue. The DME scaling method would then be appropriate.

The use of DME has been restricted to psychophysical studies for a very long time. But since Sorace's (1996) study on grammaticality judgements, the procedure has received more attention in linguistics.

Procedure:

1. Judges are presented with a modulus after every 4-10 stimuli.
2. The experimenter can either assign a fixed value such as 100 to the modulus or ask the judges to assign a value to it.
3. The judges are asked to assign a number to each of the stimuli, *relative* to the modulus. For example, if the current stimulus is twice as friendly-sounding as the modulus, it should be assigned the value 200; if it is half as friendly-sounding as the modulus, it should be assigned 50.
4. Raw scores are transformed before being subjected to statistical analyses. There are two procedures: (a) transforming raw DME scores to logarithmic scores, or (b) dividing each raw DME score by the score assigned to the modulus and then transforming this score to the logarithmic score (Lodge, 1981; Sorace, 2003).

Advantages:

1. DME does not restrict the number of values that can be used.
2. There is a great amount of evidence (though not in intonational studies) that DME yields interval data.

Disadvantages:

1. Judges do not find it an easy task to do the 'mental calculation' ('2 times more, or 2 times less') within the time allowed for making a judgement.
2. Extra procedures need to be taken to validate the use of DME.

Recommendations:

1. Make sure that judges understand how to perform magnitude estimations by including a control condition. In the control condition, judges can be asked to perform magnitude estimations of the length of a line for example.
2. State explicitly that judges should not use a restricted range of numbers (Sorace, 2003).

Examples of applications:

The DME has not been applied in previous studies of intonation. An example of applications of this scale is the present investigation (Section 5), in which we studied the perception of the meaning 'friendly' as signalled by pitch register and the perception of levels of pitch register.

3.1.4 Visual Analogue Scales (VAS)

The Visual Analogue Scale measures the intensity or magnitude of sensations and subjective feelings (e.g. pain and mood), and the relative strength of attitudes and opinions about specific stimuli. The scale used by the participant is a straight line (usually 100 mm long) with verbal descriptors (unipolar or bipolar) at each end. There are two different types of VAS, the horizontal VAS and the vertical VAS. The horizontal VAS (see the example in Figure 7) is preferred over the vertical VAS (Sriwatanakul et al., 1983), because it yields a more uniform distribution of scores (Wewers and Lowe, 1990). As for gradations on the line, it has been shown that using gradations may reduce its sensitivity. Lines shorter than 100 mm tend to result in a greater error variance (Revill et al., 1976).

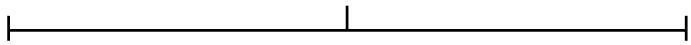

Figure 7: An example of a Visual Analogue Scale

Procedure:

1. Construct a straight horizontal line of a specified length (preferably 100 mm) with verbal descriptors at each end; the descriptors are short phrases that describe the variable to be measured and should be easily understood.
2. Present a standard stimulus which functions as an anchor at an equal interval; the consistency of the judgement process can thus be enhanced.
3. Ask the judges to put a mark on the line that best corresponds with the extent to which a specified attribute is perceived for the stimulus.
4. Measure the distance in millimetres from the mark to the left end, which is, by definition, 0 for each VAS.

Advantages:

1. The VAS is simple to use and can be administered easily.
2. It can capture subtle differences between stimuli well.

Disadvantages:

1. The test using the VAS is usually administered as a paper and pencil test, which is time-consuming to process because it involves manual scoring; computer assisted testing is thus recommended.
2. Score transformations can be misleading (Maxwell, 1978), although different kinds of transformations have been recommended: Log transformation (e.g. Bond and Lader, 1974), arcsine of the square root transformation (e.g. Snedecor and Cochran, 1967).

Recommendations:

1. If more than one attribute is to be measured for the same set of stimuli, conduct separate sessions for each attribute.
2. Make sure that judges draw the slash *on* the scale (Figure 8a), and *not above* the scale (Figure 8c) because the latter can result in scores (see the extension of the solid slash mark in Figure 8c) that are noticeably different from the scores intended (see the crossing between the solid slash and the VAS in Figure 8b).

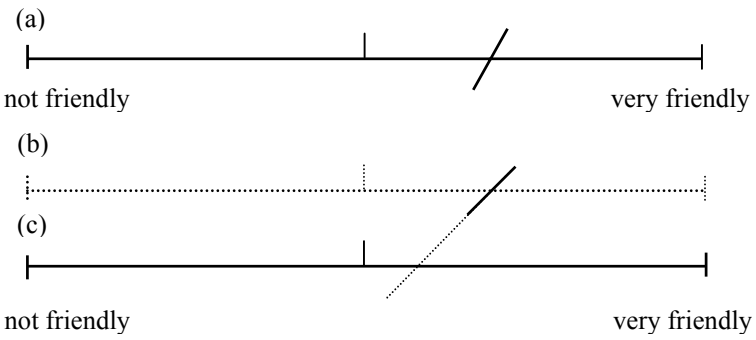

Figure 8: Examples of the appropriate way (a) and the inappropriate way of drawing the slash on a Visual Analogue Scale (c)

Examples of applications:

Although less common than the EAI scale, the VAS has been used to assess the perceived degree of a certain meaning in intonational research (Rietveld, Gussenhoven, Wichmann, and Grabe, 1999; Chen, Gussenhoven, and Rietveld, 2004; Chen, to appear). The two ends of the VAS are sometimes labelled with the adjective and its negation or antonym (e.g. friendly vs. unfriendly, patient vs. impatient, tense vs. relaxed in Rietveld et al. 1999) and sometimes labelled with the adjective and its negation with the negator 'not' (e.g. friendly vs. not friendly, confident vs. not confident, surprised vs. not surprised, emphatic vs. not emphatic in Chen et al. 2004). The possible negative influence of labelling the two ends of the scale with an adjective and its negation on judges

may not be as big as Grabe et al. (1997) pointed out for the EAI scale (see Section 3.1.1), because judges are explicitly instructed to draw a slash on the scale to indicate the perceived degree of a certain meaning rather than assigning a score relative to the two ends.

The VAS has also been used to assess perceived nonsemantic intonational properties. For example, Gussenhoven and Rietveld (1998) asked native speakers of Dutch to judge the prominence of accented syllables in utterances realised with an H*L pitch accent (see Figure 9) as an effect of peak height and of the gender of the speaker.[5] To evaluate the effect of peak height covarying with the baseline (i.e. register), separate ANOVAs were carried out on the scores for the 'female' and 'male' voices. To evaluate the effect of gender, an ANOVA was carried out on the perceived prominence values of the stimuli with the mid baseline, which were available in both the 'female' voice and the 'male' voice.

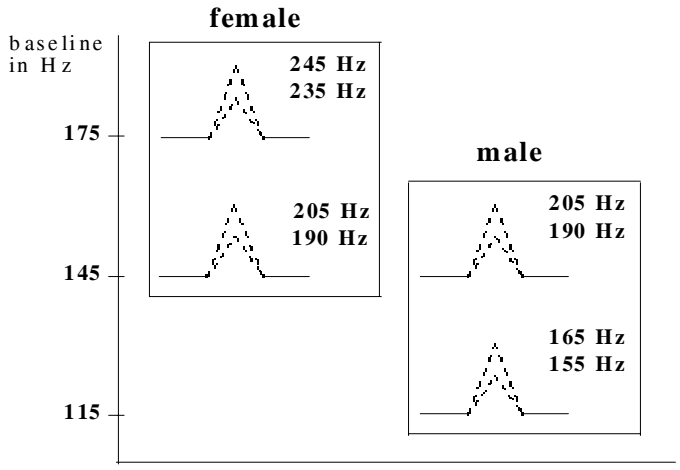

Figure 9: Structure of the experimental contours with hypothetical male and female reference scales, indicated by the boxes. Reproduced from Figure 1 of Gussenhoven and Rietveld (1998).

3.2 Multidimensional Scaling (MDS)

When people are asked to assess the similarity of two voices, they pay attention to a number of attributes: pitch, rate of speech, nasality, harshness, etc. This means that the similarity score they may be asked to give is based on a number of underlying dimensions. This is why we often need more than one

[5] In the original paper, the authors called the scale a magnitude estimation scale.

dimension to represent the perceived similarities. For example, voices A, B, and C are rated for dissimilarity (δ_{ij}): higher values δ_{ij} correspond with higher values of perceived dissimilarity. Assuming that the scores are $\delta_{AB} = 4$, $\delta_{AC} = 3$ and $\delta_{BC} = 5$, we can see that it is not possible to project these dissimilarities on one single dimension and more dimensions are required, as shown in Figure 10.

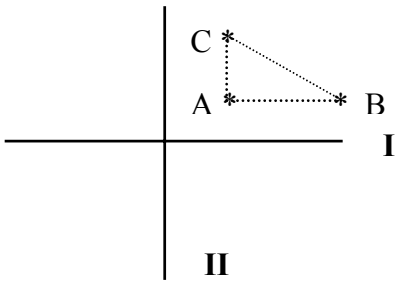

Figure 10: Objects A, B, and C projected in a two-dimensional space

The distances between A, B, and C, which are operationalised as the lengths of the lines between them, are called "Euclidean distances". They reflect much better – here even perfectly – the perceived dissimilarities.

If we know the physical dimensions, we can interpret and label the perceptual dimensions needed to represent the stimuli. In our example, the difference in speech rate between A and B is very small and between A and C quite salient, axis II can thus be labelled as the rate of speech dimension.

There are a number of possible applications of dissimilarity scaling. One is exemplified above, another is an application in which perceived dissimilarities are compared with theoretical dissimilarities, calculated on the basis of a specific metric such as the number of starred tones, or the difference between the number of H and L tones, etc. (cf. Gussenhoven and Rietveld, 1991).

A procedure that does not involve (dis)similarity judgements but does imply the use of multiple scales in the experiment itself is the one in which a factor analysis is used. The procedure boils down to a statistical inspection of the correlation coefficients calculated between the scores on the scales. If two or more scales are highly interrelated, a new underlying scale (called 'factor') is constructed on the basis of these interrelated scales. Often groups of correlated scales can be found which yield, in turn, new underlying variables. Thus a more concise description of the rated objects can be obtained. This technique was applied as early as in 1964 by Uldall.

Procedure:

1. Make pairs of all stimuli. This results in $n \times (n - 1)$ stimulus pairs.
2. Present all pairs of stimuli to the judges, preferably in different orders.
3. Ask the judges to express the dissimilarity between the two members of each pair in a number, for instance 1 to 10.
4. Apply a multidimensional scaling technique (available in SPSS and many other statistical packages).
5. Try to minimise the number of dimensions (i.e. the dimensionality of the resulting space) needed to reflect the dissimilarities in an optimal way. The degree to which the dissimilarities and the distances arising in the space created in this way agree is often expressed in a goodness-of-fit index called 'stress'.

Disadvantages:

1. The number of pairs of stimuli to be presented is quite large: $n \times (n - 1)$.
2. The determination of the optimal combination of number of dimensions and stress index is not automatic, but requires human judgement.

Recommendations:

Before starting the real experiment, the experimenter should present judges with a number of stimulus pairs in which the two members represent two extremes. In this way, judges can get an idea of the range of possible dissimilarities.

Examples of applications:

Uldall (1964) carried out an experiment with utterances spoken with synthetic intonation contours. These utterances covered variations in pitch range, boundary tone, and contour shape. Participants were asked to rate the utterances on 14 bipolar scales (e.g. 'bored – interested', 'impatient – patient', etc.; each scale comprised 7 possible values). A factor analysis was carried out on the scores obtained on the 14 scales. The purpose of the experiment was twofold: (a) to find out the relations between the rating scales, and (b) to locate the contours in the space made up by the resulting factors. Three factors were found: pleasant/unpleasant, authoritative/submissive and strong/weak. The locations of the contours along the three dimensions can be found in Uldall (1964, p. 279).

4 Criticism of the processing of data in scaling experiments: Ordinal or interval statistics?

Since Siegel's book 'Nonparametric Statistics' (1956), a large number of social scientists and researchers who tend to follow trends in social sciences, such as sociolinguists, psycholinguists, and speech/language pathologists, refrain from using interval statistics or parametric tests such as *t*-tests and *F*-ratios in ANOVA when the data at issue are strictly speaking not measured at the interval level. They assume that the statistical result of a *t*-test is affected by the mechanism that produced the numbers. Their logic is that interval statistics use the same properties of the numbers used, as the latter are assumed to reflect the properties of the objects: "7 - 6" = "3 - 2".

Many statisticians and methodologists do not follow this strict line. In the debate on this topic (many articles on this subject were published in the sixties and seventies of the 20th century) they say that "statistics does not know what's behind the numbers", and prefer to distinguish between numbers seen as scores that are assumed to reflect magnitudes of properties of objects, and numbers seen as the input of a statistical procedure designed to find out whether the differences in the numbers obtained from two (or more) groups are due to chance or not. Harris (1975, pp. 226-227) for instance, wrote that "the validity of statistical conclusions depends only on whether the numbers to which they are applied meet the distributional assumptions to derive them, and not the scaling procedures used to obtain the numbers." Incidentally, *t*-tests and *F*-ratios are known to be quite insensitive to the nonnormality of the distributions. Harris also drew attention to the fact that most scales which are assumed to be of the ordinal level – at least by some strict interpreters of Stevens' laws (Stevens, 1951) – contain interval information. To provide one example: assume we obtained the scores 3 - 4 - 7 - 9 on a scale with 10 categories. If these values are ordinal data, any monotonic transformation is permissible (see Section 3), which means that a transformation resulting in the values 3 - 4 - 7 - 999 would be equally valid. This is obviously not the case, and apparently the intervals between the numbers used contain some information.

This led both Harris (1975) and Labovitz (1972, p. 515) to rather categorical statements on this matter. We quote the latter: "Empirical evidence support the treatment of ordinal variables as if they conform to interval variables (.....). Although some small error may accompany the treatments of ordinal variables as interval, this is offset by the use of more powerful, more sensitive, better developed, and more clearly interpretable statistics with known sampling error." On the basis of these statements – and many more similar statements can be quoted (e.g. Anderson, 1961) – we think that ANOVA can be applied to the data that are not strictly of the interval level, but contain information on the magnitude of the differences between objects, as obtained in magnitude scaling and equal appearing interval scaling. Munro and Page (1993) state that "there

is now sufficient evidence to show that use of parametric tests with ordinal data rarely distorts the results". Dexter and Chestnut (1995) carried out an interesting simulation experiment, in order to find out whether the analysis of VAS data with parametric and non-parametric tests yielded different results. Their conclusion is that "*t* and ANOVA are good choices to compare VAS measurements among groups". In spite of the arguments given above, users of parametric statistics should remain careful and verify whether assumptions like homogeneity of variances are met, especially in the case of small numbers of observations (Rietveld and van Hout, 2005).

5 Present investigation

Given the advantages and disadvantages of each scale, it is a complex task to determine which scales are most suitable for obtaining perceptual judgements in intonation research. In this study, we will evaluate three scales, the Direct Magnitude Estimation Measurement, the Equal Appearing Interval Scale, and the Visual Analogue Scale, on the basis of data obtained from Dutch listeners who rated the degree of friendliness signalled by pitch register in Dutch (Experiment 1) and data from a different group of Dutch listeners who listened to the same stimulus set but rated the level of pitch register (Experiment 2). In the light of the evidence provided in Section 4, we will use parametric statistics to analyse data obtained from the present study. In the interpretation of the results, the magnitudes of the observed effect sizes (i.e. measures of the degree of association between the effect, which can be a main effect, an interaction, or a linear contrast, and the dependent variable) play an important role, in addition to the presence of significance. Significance can easily be obtained when the degrees of freedom associated with the error term are high, as is the case with our ANOVAs. Consequently, a *p*-value is not a good index of the relevance of an effect. We thus pay great attention to effect sizes of the variables in our interpretation of the results. The measure of effect size we adopted here is partial eta squared ($\eta^2_p = SS_{factor}/(SS_{factor} + SS_{error})$). For example, if a *p*-value of 0.031 is obtained ('significant at the 5% level'), but η^2_p is only 0.14, we consider the relevance of the significant variable low. This approach is in line with recent developments in statistical reporting, in which researchers are advised not to focus solely on *p*-values (cf. Krantz, 1999).

5.1 Experiment 1

In a previous study, Chen et al. (2004) examined the perception of friendliness as signalled by pitch register in British English and Dutch by means of the Visual Analogue Scale. It was found that native speakers of British English and Dutch differed in their ratings in two ways. First, although by and large, an

increase in register led to an increase in the perceived degree of friendliness, as predicted by Ohala's (1983; 1984) Frequency Code, the increase was steeper in English listeners' ratings than in Dutch listeners' ratings. Second, in Dutch listeners' ratings, there was an observable decrease in the perceived degree of friendliness at the highest level of register adopted in their stimuli.

In Experiment 1 to be reported here, native speakers of Dutch listened to a subset of Chen et al's Dutch stimuli and rated the degree of friendliness for each stimulus on three scales, the DME with 100 as the modulus value, the EAI with 7 points, and the VAS (100 mm), which was presented horizontally.

5.1.1 Experimental design

Sixty stimuli were selected from Chen et al.'s (2004) Dutch stimulus set. Their stimuli were generated from source utterances recorded by a female British English-Dutch bilingual speaker using a sampling rate of 48 kHz. The recording was subsequently down sampled to 32 kHz and subjected to speech manipulation (i.e. removing the original pitch and imposing new pitch patterns), which was performed by means of Praat (Boersma and Weenink, 1996). The stimuli selected for this study can be grouped under three independent variables, Pitch Register (5 levels), Pitch Contour (2 levels: H*L L%, L*H H%), and Speech Act (2 levels, Information, Instruction). Each speech act was realised on three utterances. Pitch contour and Speech Act were included because they were shown to interact with the interpretation of intonational meaning in Chen et al. (2004). Examples of the two speech acts in Dutch as well as their translations in English are given in (3). The accented syllables are in capitals.

(3) Information: *Wat is het niVEAU van deze cursus?*
(What's the LEvel of this course?)
Instruction: *Je moet het declaRAtie formulier invullen.*
(You should fill out the CLAIM form.)

The schematic representations of H*L L% and L*H H% are given in Figure 11. Pitch register was varied in five levels by raising both the H tones (H*, H, H%) and the L tones (%L, L*, L, L%) in four equal steps of 20Hz.

Test tapes were prepared containing experimental stimuli and practice trials. A 4.5s pause was inserted after each stimulus to allow participants to give their judgements. There was a 7s pause between blocks of 10 stimuli and between blocks of four practice trials. Each block was preceded by a 200ms 300Hz sine wave to signal the beginning of the block. The anchor used in Chen et al. (2004) was inserted before the first item of each block. It was a neutral-sounding realisation of the utterance *Verkoopt u ook bioLOgisch fruit?* (*Do you sell orGAnic fruit as well?*) with the pitch contour %L H*L H%. The an-

chor was included to provide listeners with a reference point for their judgements on the semantic scales. It is known that the distance between a given stimulus and the anchor can have an effect on the perceptual judgement. To minimise this effect, the stimuli were randomised such that those representing the same experimental condition could appear at various places in a stimulus block. Each of the stimulus orders was recorded onto DAT tape (48 kHz at 16 bits) and then copied to a TDK audio tape. This led to two 8-minute test tapes.

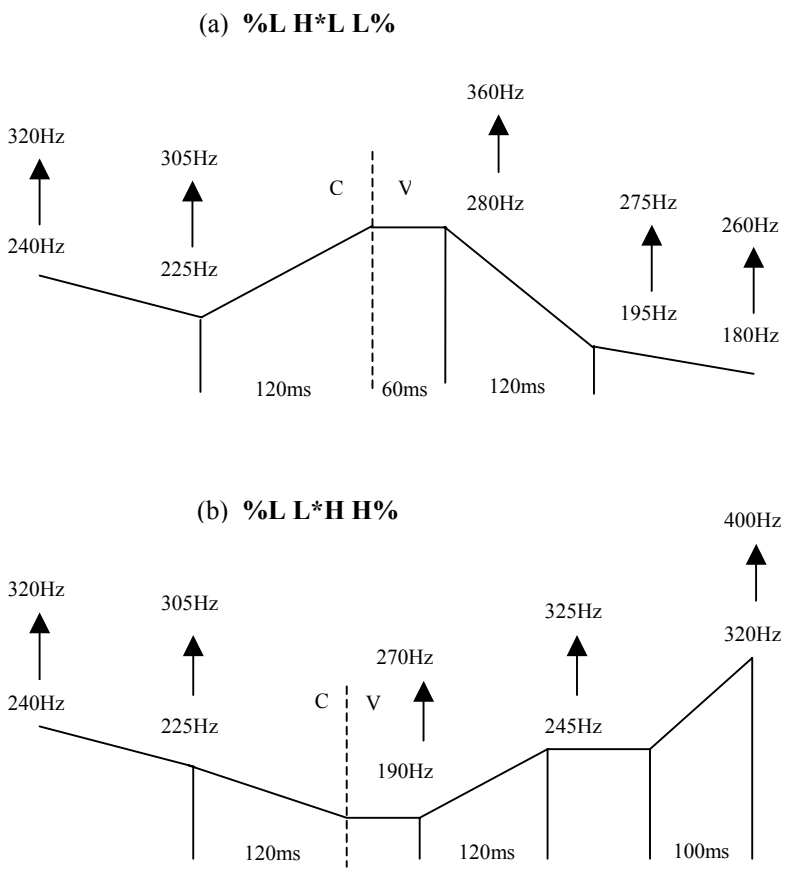

Figure 11: H*L L% and L*H H% with pitch register varied within a range of 80Hz. Reproduced from Figure 4 of Chen, Gussenhoven, and Rietveld (2004).

5.1.2 Procedure

Twenty-four native speakers of Dutch recruited from PhD students at the Radboud University Nijmegen took part in this experiment. All participants reported having normal hearing. The experiment was carried out in three sessions, in each of which one of the three scales was used. As our design was created for within-subject testing, an interval of at least 48 hours was required between two consecutive sessions in order to minimise any learning effect. Moreover, participants were randomly assigned to six groups. Each group received the three scales in a different order. Within each group, two participants were randomly assigned to each of the two stimulus orders.

Participants were not briefed on the purpose of the experiment. They were told that the experimenters were interested in the signalling of friendliness in Dutch. They were instructed by means of written instructions to try to imagine themselves as the addressees of the stimuli and indicate for each stimulus how friendly the speaker sounded.

We presented the stimuli through a Philips AQ6455 cassette recorder/player at an adequate volume in a quiet room. Participants were tested individually. Each session took about 10 minutes.

5.1.3 Statistical analyses and results

Three sets of data containing 'friendliness' scores were obtained from the three scales. Scores from the DME were also transformed into logarithmic scores (Lodge, 1981; Sorace, 2003). This gave us the fourth set of data. One repeated measures ANOVA was performed on each data set with three factors, Pitch Contour, Speech Act and Pitch Register, at a significance level of 0.05. For the data of each participant, the mean score of the stimuli representing the same experimental condition was calculated by dividing the sum of the scores by the total number of stimuli. As regards missing values (11 in total), the mean was estimated by dividing the sum of the scores by the total number of stimuli minus the number of stimuli that were not rated. In repeated measures designs, an extra assumption has to be met, i.e. the assumption of sphericity. This assumption boils down to the condition that the variances of all differences between levels of the within-subject factors are equal. If the condition is not met, a number of corrections based on estimates of sphericity can be applied to the degrees of freedom used to access the observed *F*-ratio. These corrections include the Greenhouse-Geisser correction and the Huynh-Feldt correction. We used the Huynh-Feldt correction because the Greenhouse-Geisser correction is too conservative (Field, 2003). An overview of the results of the analyses is given in Table 1, where the η^2_p values are reported and the significance is based on *F*-ratios.

As we can see, the results are similar across the analyses with the exceptions of the effect of Speech Act and the two-way interaction of Contour and Speech Act. Specifically, there was a main effect of Speech Act in the analysis with logarithmic DME scores, as in the analyses with EAI and VAS scores, but not in the analysis with raw DME scores. The interaction of Contour and Speech Act reaches significance in the analyses with DME log scores, DME scores and EAI scores but not in the analysis with VAS scores.

	Contour (C)	Speech act (SA)	Register (R)	C × SA	C × R	SA × R	C × SA × R
DME	*	n.s	*	*	n.s	n.s	n.s
	η^2_p = .377	η^2_p = .070	η^2_p = .418	η^2_p = .228			
DME	*	*	*	*	n.s	n.s	n.s
(log)	η^2_p = .338	η^2_p = .189	η^2_p = .420	η^2_p = .377			
EAI	*	*	*	*	n.s	n.s	n.s
	η^2_p = .438	η^2_p = .231	η^2_p = .455	η^2_p = .346			
VAS	*	*	*	n.s	n.s	n.s	n.s
	η^2_p = .465	η^2_p = .275	η^2_p = .469	η^2_p = .054			

Table 1: Overview of the results from DME, DME-log, EAI and VAS in the perception of 'friendly'; the effect size is indicated by η^2_p; * marks significant effects at the .05 level.

Figure 12 illustrates the relations between Contour and Speech act in the data obtained from each scale. Common to all the data sets is that L*H H% was rated more friendly than H*L L%. In the data from EAI, DME, and DME-log, H*L L% was rated more friendly in speech act 1 (Information) than in speech act 2 (Instruction) (4.08 vs. 3.76 in the EAI data, 114.95 vs. 104.18 in the DME data, 1.99 vs. 1.90 in the DME-log data), while the rating of L*H H% differed between speech acts to a lesser degree (4.38 vs. 4.31 in the EAI data; 124.01 vs. 125.83 in the DME data, 2.024 vs. 2.020 in the DME-log data). In contrast, in the data from VAS, the rating of L*H H% differed between speech acts (53.27 vs. 51.09) to a similar degree that the rating of H*L L% did (47.08 vs. 44.13). This difference between the data from EAI, DME and DME-log on the one hand and the data from VAS on the other hand accounts for the significance of the interaction between Contour and Speech Act in the former but lack of significance in the latter.

A possible explanation for the abovementioned difference in the results is that VAS provided the listeners with a relatively large 'rating space' in comparison to EAI and the DME such that they could indicate the subtle difference in the perceived friendliness of L*H H% in different speech acts. The limited rating space of the EAI has been discussed by Stevens (1974). It is, however, not straightforward why the DME, with no upper limit, has a limited rating

space. We have observed that the participants used 33 different scores in the DME scale. 60% of the scores fell in the region from 100 to 200 and 33% of the scores fell in the region from 1 to 99. Within each region, the scores were not uniformly distributed. This suggests that in spite of the assumed abundant rating space of the DME, the participants used it very parsimoniously. This is plausible considering that the use of DME entails some mental arithmetic on the part of the listeners, who usually have less than five seconds to give their judgement for each stimulus. As a result, the listeners may not use the DME as they are supposed to. As regards the main effect of speech act, it appeared to be canceled out by the opposite trends in the scores for H*L L% and L*H H%. This can again be explained by the sparing use of the DME. As the participants employed a small number of scores most of the time, they could not reliably indicate the limited difference between speech acts for L*H H%. However, this drawback of the DME appeared to be remedied when the DME scores were converted to log scores.

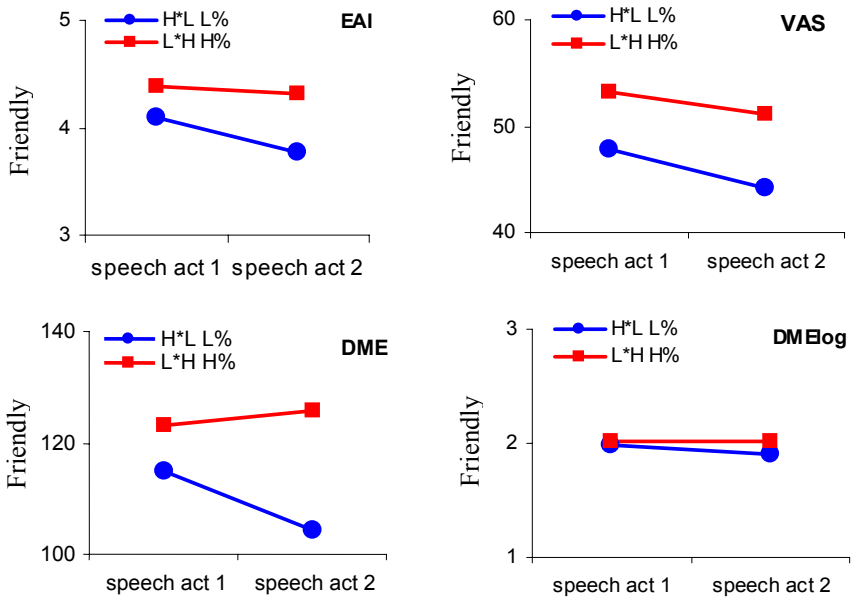

Figure 12: The effects of the two-way interaction of Contour and Speech Act in data obtained from the EAI, the VAS, the DME, and the DME-log data

To sum up, results from Experiment 1 show that the VAS is sensitive enough to capture subtle differences in the perceived meaning in different experimental conditions but the EAI scale and the DME measurement are not. Second,

when the DME measurement is used, it is desirable to convert the raw DME scores to the log scores.

5.2 Experiment 2

We have seen in Experiment 1 that, although the results were similar across the scales used to obtain perceptual judgements, the VAS appears to be more sensitive than the other scales in capturing differences in perceived intonational meaning. In Experiment 2, we addressed the question as to which of the scales can most accurately reflect listeners' perception of pitch register. Note that in Experiment 1, we do not have direct evidence or a model available that relates 'friendliness' scores to independent variables like Speech Act and Pitch Contour, except for Ohala's (1983; 1984) Frequency Code, which associates a higher friendliness score with a higher pitch register. In contrast, in Experiment 2, the independent variable Pitch Register is more directly affected by known physical variables, such as mean F0 of the whole contour, the F0 values of stretches of L-targets, etc. (cf. Rietveld and Vermillion, 2003). This would allow us to find more direct evidence for a favourable scale to obtain perceptual judgements of the physical make-up of intonation contours.

Related to the perception of pitch register is the issue of how to represent variations in register. Because the physical form of pitch is linear by nature but our perception of pitch generally is not, it is often suggested in the literature that there is a need to transform the Hz scale into a psychoacoustic scale, such as the musical semitone scale, the mel scale, and the ERB-rate scale (Equivalent Rectangular Bandwidth). The semitone scale is a logarithmic transformation of the Hertz scale ($ST = 39.87 \times \log (Hz/50)$). The mel scale is linear below 500Hz but logarithmic above ($mel = 2595 \times \log (1+ Hz/700)$). The ERB-rate scale is between linear and logarithmic below 500Hz but logarithmic above 500Hz ($ERB = 16.6 \times \log (1 + Hz/165.4)$).

In order to establish which scale should be used for intonation, Nolan (2003) asked participants to replicate pitch contours produced by a male and a female speaker in three pitch spans (neutral, compressed, and expanded). The goodness of fit of the pitch spans of the imitations was evaluated when pitch was presented in each of the four scales. Results for both female and male participants show that the semitones and ERB-rate best reflect the participants' intuition about intonational equivalence (i.e. span), with semitones marginally better. In Experiment 2, we wanted to find out whether it is necessary to represent pitch register in a psychoacoustic scale.

To these ends, we examined the perception of pitch register represented in Hertz, semitones, mel and ERB-rate on each of the three scales, DME (with 100 as the modulus value), EAI (with 7 points), and VAS.

5.2.1 Method

The stimulus tapes made for Experiment 1 were used. The instructions used in Experiment 1 were adjusted such that the task was to indicate the register of each stimulus on a scale.

5.2.2 Procedure

A different group of 24 native Dutch speakers were recruited from the PhD students at the Radboud University Nijmegen and student assistants at the Max Planck Institute for Psycholinguistics. The experiment was conducted following the same procedure as in Experiment 1.

5.2.3 Statistical analyses and results

Four sets of data containing 'register' scores were obtained from the four scales. As in Experiment 1, scores from the DME were transformed to logarithmic scores. This gave us the fourth set of data. One repeated measures ANOVA was performed on each data set with three factors, Pitch Contour, Speech Act and Pitch Register. Missing values (4 in total) were treated in the same way as in Experiment 1. Table 2 presents an overview of the results of the analyses.

	Contour (C)	Speech act (SA)	Register (R)	C × SA	C × R	SA × R	C × SA × R
DME	n.s	n.s	*	n.s	*	n.s	*
	η^2_p = .027	η^2_p = .000	η^2_p = .661	η^2_p = .131	η^2_p = .154	η^2_p = .04	η^2_p = .132
DME	n.s	n.s	*	n.s	*	n.s	*
(log)	η^2_p = .104	η^2_p = .726	η^2_p = .731	η^2_p = .11	η^2_p = .199	η^2_p = .019	η^2_p = .117
EAI	n.s	n.s	*	n.s	*	*	n.s
	η^2_p = .000	η^2_p = .002	η^2_p = .921	η^2_p = .002	η^2_p = .147	η^2_p = .188	η^2_p = .068
VAS	n.s	n.s	*	*	n.s	n.s	*
	η^2_p = .015	η^2_p = .011	η^2_p = .754	η^2_p = .380	η^2_p = .05	η^2_p = .081	η^2_p = .14

Table 2: Overview of the results from DME, DME-log, EAI and VAS in the perception of 'Register'; the effect size is indicated by η^2_p;* marks significant effects at the .05 level.

The results obtained with the four scales for the main effect 'Register' of Experiment 2 are quite similar and significant, with large effect sizes, especially for EAI (η^2_p = 0.921). Some discrepancies can, however, be observed in the interactions. Whereas, for instance, the C×R interaction on the VAS scale was not significant, it was significant on the DME, DME-log, and EAI scales, al-

though the observed effect sizes were quite small (around 0.15). Intriguingly, two significant interactions were found in data obtained from each scale but they were not the same interactions. Is there an explanation for these differences?

Three factors may be relevant to account for the presence or absence of a significant interaction. First of all, if an interaction is highly predictable from the model that relates the dependent variables to the independent variables, it is very likely that it will be found to be significant. In Experiment 2, we have reason to expect a significant interaction between Contour and Register in the perception of pitch register. This two-way interaction was indeed found to be significant in data obtained by means of the EAI, the DME, and the DME-log. On this ground, one may tentatively posit that the VAS is less suitable to measure the perception of the physical make-up of intonation, such as pitch register. On the other hand, we have no specific reason to expect a significant interaction between Contour and Speech Act in the perception of pitch register.

The second factor is related to the rating space that participants are actually able to make use of (as opposed to the assumed rating space). As mentioned in Section 5.1, participants can be constrained by a limited rating space. Consequently, they fail to indicate subtler differences between levels of independent variable B (e.g. Speech Act) at one level of independent variable A (e.g. Contour), while being able to indicate greater differences in perception between levels of independent variable B at another level of independent variable A. This would lead to a significant interaction between independent variables A and B that would otherwise be insignificant. This may account for the significant two-way interaction between Speech Act and Register found in the data obtained by the EAI scale and the three-way interaction between Speech Act, Contour and Register found in data obtained from the DME and the DME-log.

The third factor relates to the error terms, which directly affect *F*-ratios and the associated *p*-values, in addition to degrees of freedom (identical across tests using different scales in our experiment) and effect sizes (overall relatively small here). The error term used in testing the effects is made up of the four-way interaction between the factor Participants and the within-subject factors, in our case Register, Contour and Speech act. It might get a high value if the judgements of the participants differ as a function of the levels of the within-subject factors. If the value of this interaction term is high, a relatively small effect of the interaction between two within-subject factors will possibly not exceed the error term; as a consequence the interaction effect will not reach significance. However, we decided not to test this factor because of lack of an appropriate error term (cf. Rietveld and van Hout, 2005). Tukey's test is available to assess the interaction between participants and a within-subject factor (Tukey's test of additivity). It deals with only two-way interactions and is thus not suitable for our design involving a four-way interaction.

In order to assess differences in how participants made use of the rating space allowed by the three different scales, the coefficient of variation (COV = SD/Mean) was determined for each participant for the scores obtained for each level of the variable Register (there are 5 levels); this was done for each scale and each contour. Subsequently, SDs of the coefficients of variation were calculated, which reflect the variability of the COVs. Large SDs can be expected in scales that allow a large range of scores (such as VAS and DME), smaller SDs in scales that allow a small range of scores (such as EAI and DME-log). A large range of scores can be an indicator of a large rating space, provided that the scores are not skewed to a certain range. The COVs and SDs were pooled over the two levels of Speech Act, as presented in Table 3.

Scale	Contour	mean coefficient of variation	SD of coefficients of variation
EAI	H*L L%	.266	.077
	L*H H%	.243	.062
VAS	H*L L%	.253	.134
	L*H H%	.242	.117
DME	H*L L%	.375	.174
	L*H H%	.344	.132
DME-log	H*L L%	.088	.052
	L*H H%	.075	.033

Table 3: Coefficient of variation and the corresponding standard deviation pooled over the two levels of Speech Act for each contour in data obtained via EAI, VAS, DME and DME-log.

The results confirm our expectations: relatively high values for VAS, relatively low values for EAI and DME-log. We might therefore expect larger interactions between participants and within-subject factors for VAS and thus smaller *F*-ratios than for the other scales. This may in turn account for the absence of a significant two-way interaction between Contour and Register in the data obtained by the VAS. As for DME, although we obtained relatively high values, this result can not be taken to reflect the usage of a large rating space. Similar to the ratings in Experiment 1, participants used a large number of different scores, varying from 1 to 958. But over 70% of the scores fell in the region from 100 to 200. Within each region 100-200, the scores were not uniformly distributed, as in Experiment 1. In contrast, we did not observe such patterns in VAS scores.

To sum up, considering the criterion predictability of the significance of the interactions, the EAI scale and the DME measurement outperform the VAS, whereas the VAS outperforms the EAI scale and the DME measurements considering the criterion rating space. It is thus still not quite clear which scale is most suitable for obtaining perceptual judgements of the physical make-up of intonation. On the basis of the important practical concern that participants uniformly find the DME less than easy to use and do not use it as intended, we

suggest that DME and DME-log are less ideal than EAI and VAS. In order to compare EAI and VAS, we plotted the range of scores obtained from the EAI scale and the VAS as a function of Contour and Register, as shown in Figure 13; the scores were made comparable by a percentage transformation. As is apparent, participants distinguished a noticeably larger difference between the lowest and the highest pitch registers when making judgements on the EAI scale than when making judgements on the VAS. We know in advance that the perceptual differences between levels of the factor Register will be linear, as the perceptual differences of the physical frequency scale are quasi-linear in this range of F0 values. These differences appear to be more clearly expressed when EAI is used. We thus argue that the EAI scale is more suitable for obtaining perceptual judgements on the physical make-up of intonation, such as pitch register.

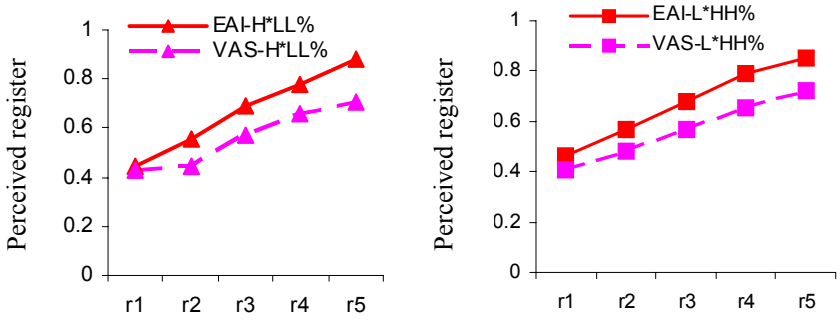

Figure 13: The range of scores obtained from the EAI scale and the VAS as a function of Register and Contour. The scores were made comparable by means of a percentage transformation.

Our second goal in Experiment 2 was to determine how well the three scales reflect listeners' perception of register represented in different physical and psychoacoustic scales. To this end, we conducted regression analyses with the mean register score obtained for each level of the variable Register (the register score hereafter) as the dependent variable, and the mean F0 for each level of the variable Register (the mean register hereafter) as the predictor. In detail, the mean register score was calculated for each level of Register per perceptual scale. The mean register was derived for each level of Register from the mean F0 values that made up the contour of each stimulus (Rietveld and Vermillion, 2003). The mean registers were then transformed from Hertz to each of the three psychoacoustic scales. Combining the mean registers in Hertz, semitones, mels and ERB-rate with the register scores obtained from DME, EAI, and VAS gave us another 12 sets of data. In Table 4, we present the R, the R^2, and the *p*-values of the correlations from each regression analysis. The correla-

tions were substantially higher when register was represented in mels and Hertz than when register was represented in semitones and ERB-rate. As the number of register values was very small, i.e. 5, the power of the analysis was very low. A non-significant correlation should therefore not be taken as evidence against the presence of a correlation in the population. Table 4, however, presents no evidence that the simple Hertz scale should be avoided when physical characteristics of contours are to be correlated with perceptual scores, especially within the same gender. Moreover, the semitone scale and the ERB-rate were not found to best reflect the variations in pitch register. This is opposite to the finding of Nolan (2003) that these two scales were better than the mel scale. The discrepancy in the findings can be accounted for by the fact that we were concerned with pitch register while Nolan (2003) was concerned with pitch span, and we operated with one level of the factor gender.

Register	Scale	R	R^2	*p*-values
Hertz	DME	.996	.993	p < .01
	EAI	.998	.996	p < .01
	VAS	.995	.990	p < .01
Semitone	DME	.847	.717	n.s
	EAI	.850	.723	n.s
	VAS	.848	.719	n.s
Mel	DME	1.000	.999	p < .01
	EAI	.994	.988	p < .01
	VAS	.996	.998	p < .01
ERB-rate	DME	.850	.722	n.s
	EAI	.853	.728	n.s
	VAS	.852	.726	n.s

Table 4: Overview of results from the regression analyses with perceived Register as the dependent variable, and the 5 levels of Register as predictor.

5.3 Conclusions

In this investigation, we evaluated the suitability of three scales, the DME measurement, the EAI scale, and the VAS, for obtaining perceptual judgements of intonational meaning (Experiment 1) and the physical make-up of intonation (Experiment 2). Identical independent variables were used in the two experiments, i.e. Pitch Contour, Pitch Register and Speech Act. Results from Experiment 1 suggest that the VAS is most suitable for obtaining perceptual judgements of semantic properties of intonation. The DME measurement and EAI scale appear to provide judges with limited rating space and can lead to significant interactions between independent variables that are otherwise not significant. In contrast, Experiment 2, which dealt with the perception of the physical make-up of intonation, shows that the EAI scale outperforms the

other scales. In our stimuli, five levels of pitch register were included, while the EAI scale had 7 levels. We suggest that it is essential to include at least the same number of levels into the EAI as the number of levels that the pitch related independent variable has. In case that the pitch-related independent variable does not have predefined levels, a four-point EAI scale may be sufficient (Jensen and Tøndering, 2005). The different outcomes of the two experiments can be related to the nature of the contrasts under investigation. When dealing with the perception of semantic differences, the researcher cannot decide a priori how many levels of meaning can be distinguished for a given attribute. It is therefore reasonable to use a scale that is not divided into visible intervals, like the VAS.[6] However, when dealing with the perception of pitch-related properties, the researcher often knows in advance how many levels the pitch-related variable has and can adjust the number of levels of the EAI accordingly.

References

Allard, F. (2001): Information processing in human perceptual motor performance. Course notes, University of Waterloo.

Anderson N. H. (1961): Scales and statistics: Parametric and nonparametric. Psychological Bulletin 58, 305-316.

Boersma, P. and D. Weenink (1996): PRAAT: A system for doing phonetics by computer. Report of the institute of Phonetic Sciences of the University of Amsterdam 132. (http://fonsg3.let.uva.nl/paul/praat.html)

Bolinger, D. (1965): Forms of English. Cambridge, MA: Harvard University Press.

Bond, A. and M. H. Lader (1974): The use of analogue scales in rating subjective feelings. British Journal of Medical Psychology 47, 211-218.

Caspers, J. (2000): Experiments on the meaning of four types of single-accent intonation patterns in Dutch. Language and Speech 43(2), 127-161.

Chen, A., C. Gussenhoven, and T. Rietveld (2004): Language-specificity in the perception of paralinguistic intonational meaning. Language and Speech 47(4), 311-349.

Chen, A. (2005): Universal and language-specific perception of paralinguistic intonational meaning. PhD dissertation, Radboud University Nijmegen. Utrecht: LOT.

Chen, A. (to appear): Language-specificity in the perception of continuation intonation. In C. Gussenhoven and T. Riad (eds.): Tones and Tunes: Studies in Word and Sentence Prosody. Berlin: Mouton de Gruyter.

[6] The software 'Toolist' has been developed by the Department of Linguistics for the Department of Remedial Education at the Radboud University Nijmegen, which makes it possible to present audio and visual stimuli online and automatically obtain listeners' VAS ratings by means of the computer mouse.

Cruttenden, A. (1997): Intonation (2nd ed.). Cambridge: Cambridge University Press.Dexter, F. and D. Chestnut (1995): Analysis of statistical tests to compare Visual Analogue Scale measurements among groups. Anesthesiology 82(4), 896-902.

D'Imperio, M. and D. House (1997): Perception of questions and statements in Neapolitan Italian. Proceedings of Eurospeech, Rhodes, Greece, vol. 1, 251-254.

Eggen, J. H. (1992): On the quality of synthetic speech: Evaluation and Improvements. Unpublished PhD thesis, University of Eindhoven.

Field, A. (2003): Discovering statistics using SPSS for Windows. London: SAGE Publications.

Gescheider, G. A. (1976): Psychophysics, method and theory. Hillsdale, NJ: Lawrence Erlbaum.

Grabe, E., C. Gussenhoven, J. Haan, E. Marsi, and B. Post (1997): Preaccentual pitch and speaker attitude in Dutch. Language and Speech 41(1), 63-85.

Gussenhoven, C. (1984): On the grammar and semantics of sentence accents. Dordrecht: Foris.

Gussenhoven, C. (1999): Discreteness and gradience in intonational contrasts. Language and Speech 42, 283-305.

Gussenhoven, C. (2002): Intonation and interpretation: Phonetics and phonology. In: B. Bel and I. Marlien (eds.): Proceedings of the Speech Prosody, Aix-en-Provence, 47-57.

Gussenhoven, C. and T. Rietveld (1991): An experimental evaluation of two nuclear-tone taxonomies. Linguistics 29, 423-449.

Gussenhoven, C. and T. Rietveld (1998): On the speaker-dependence of the perceived prominence of F0 peaks. Journal of Phonetics 26 (4), 371-380.

Halliday, M. K. K. (1967): Intonation and grammar in British English. The Hague: Mouton.

Harris, R. J. (1975): A primer of multivariate statistics. London: Academic Press.

Jensen, C. and J. Tøndering (2005): Choosing a scale for measuring perceived prominence. In: Proceedings of Interspeech, Lisbon, Portugal, 2385-2388.

Keijsper, C. (1984): Vorm en betekenis in Nederlandse toonhoogtecontouren, parts 1 and 2. Forum der Letteren 25, 20-37, 113-126.

Kohler, K. (1987): Categorical pitch perception. Proceedings of the 11th International Congress of Phonetic Sciences, 331-333.

Krantz, D. H. (1999): The null hypothesis controversy in psychology. Journal of the American Statistical Association 44, 1372-1381.

Labovitz, S. (1972): The assignment of numbers to rank order categories. American Sociological Review 35, 515-524.

Ladd, D.R. (1978): The structure of intonational meaning. PhD thesis, Cornell University.

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

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

Liberman, M. and I. Sag (1974): Prosodic form and discourse function. Proceedings of the Chicago Linguistics Society 10, 416-427.

Likert, R. (1932): A Technique for the Measurement of Attitudes. New York: McGraw-Hill.

Lodge, M. 1981. Magnitude Scaling: Quantitative Measurement of Opinions. Beverley Hills, CA: Sage Publications.

Maxwell, C. (1978): Sensitivity and accuracy of the Visual Analogue Scale: A psychophysical classroom experiment. British Journal of Clinical Pharmacology 6, 15-24.

Munro, B. and E. Page (1993): Statistical Methods for Health Care Research. Philadelphia: J.B. Lippincott.

Nolan, F. (2003): Intonational equivalence: An experimental evaluation of pitch scales. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 771-774.

O'Connor, J. D. and G. F. Arnold (1973): Intonation of colloquial English. London: Longman.

Ohala, J. (1983): Cross-language use of pitch: An ethological view. Phonetica 40, 1-18.

Ohala, J. (1984): An ethological perspective on common cross-language utilization of F0 in voice. Phonetica 41, 1-16.

Pierrehumbert, J. B. and J. Hirschberg (1990): The Meaning of Intonational Contours in the Interpretation of Discourse. In: R. R. Cohen, J. Morgen, and M. E. Pollack (eds.): Intentions in Communication. Boston, MA: MIT Press, 271-311.

Pike, K. L. (1945): The intonation of American English. Ann Arbor: University of Michigan Press.

Revill, S. I., J. O. Robinson, M. Rosen, and M. I. J. Hogg (1976): The reliability of a linear analogue for evaluating pain. Anaesthesia 31, 1191-1198.

Rietveld, T., C. Gussenhoven, A. Wichmann, and E. Grabe (1999): Communicative effects of rising and falling pitch accents in British English and Dutch. In: Proceedings of the ESCA workshop on dialogue and prosody, 111-116.

Rietveld, T., J. Haan, L. Heijmans, and C. Gussenhoven (2002): Explaining attitudinal ratings of Dutch rising contours: Morphological structure vs. the Frequency Code. Phonetica 59, 180-194.

Rietveld, T. and P. Vermillion. (2003): Cues for perceived pitch register. Phonetica 60, 261-272.

Rietveld, T. and R. van Hout (2005): Statistics for language research: Analysis of variance. Berlin: Mouton de Gruyter.

Sag, I. and M. Liberman (1975): The intonational disambiguation of indirect speech acts. Proceedings of the Chicago Linguistics Society 11, 487-497.

Scheffé, H. (1952): An analysis of variance for paired comparisons. Journal of the American Statistical Association 47, 381-400.

Siegel, S. (1956): Nonparametric Statistics. New York: McGraw-Hill.

Snedecor, G. W. and W. G. Cochran (1967): Statistical Methods. 6th Ed. Iowa: Ames.

Sorace, A. (1996): The use of acceptability judgements in second language research. In: V. T. Bhatia and W. Ritchie (eds.): Handbook of Second Language Acquisition. New York: Academic Press, 375-409.

Sorace (2003): Magnitude estimation of linguistic acceptability: Applications to research on developing grammars. http://www.let.uu.nl/~Sharon.Unsworth/personal/Sorace.ppt

Sriwatanakul, K., W. Kelvie, L. Lasagna, J. F. Calimlim, O. F. Weis, and G. Mehta (1983): Studies with different types of visual analogue scales for measurement pain. Clinical Pharmacology Therapy 34(2), 234-239.

Stevens, S. S. (1951): Handbook of Experimental Psychology. New York: John Wiley.

Stevens, S. S. and E. H. Galanter (1957): Ratio scales and category scales for a dozen perceptual continua. Journal of Experimental Psychology 54, 377-411.

Stevens, S. S. (1974): Perceptual magnitude and its measurement. In: E. C. Carterette and M. P. Friedman (eds.): Handbook of perception. Vol. II: Psychophysical judgement and measurement. New York: Academic Press, 361-389.

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

Thurstone, L. L. (1928): Attitudes can be measured. American Journal of Sociology 33, 529-554.

Uldall, E. (1964): Dimensions of meaning in intonation. In: D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, and J. L. M. Trim (eds.): In Honour of Daniel Jones: Papers contributed on the Occasion of his Eightieth Birthday, 12 September 1961. London: Longman, 271-279.

Wewers, M. E. and Lowe N. K. (1990): A critical review of visual analogue scales in the measurement of clinical phenomena. Research in Nursing and Health 13, 227-236.

Wichmann, A. (1991): Falls and perceptual effects. In: Proceedings of the 12th International Congress of Phonetic Sciences, 194-197.