A NEW METHOD IN INTONATION RESEARCH USING PARTLY CONTROLLED, SIMULATED DIALOGUES

I. Hertrich (1) & R. D. Gartenberg (2)

(1) Institut für Phonetik und digitale Sprachverarbeitung, Univ. Kiel
(2) EUROLOGOS S.A., Bruxelles

Abstract

Adequate speech recordings for intonation research should meet the following conditions:
(1) Naturalness
(2) Control of semantic, linguistic and pragmatic content
(3) Standardization of the recording situation.

A method is described here using simulated dialogues comprising two alternating roles: Role one serves to set the desired situational framework for subsequent responses and is prerecorded on tape. Role two consists of fixed responses presented on a monitor screen at the appropriate moment in the dialogue. A subject is requested to play the 'life' dialogue partner, listening to role one and using the presented responses. His/her spoken utterances are recorded on tape for subsequent analysis.

Some possible applications are discussed.

2. Background to the need for a new method

In a research project at the University of Kiel, a peak pattern model has been developed to describe the intonation of short, German, declarative utterances containing a single pitch-accent. Initial efforts to expand the model, applying it to utterances containing two pitch accents, had encountered difficulties in the elicitation of natural, consistent, controlled production data from naive speakers.

The study of terminal single-peaked utterances had not required a special method for recording production data because the range of variation in such utterances is much more limited, and because clear-cut hypotheses existed about the 3 possible intonational categories ('early peak', 'medial peak' and 'late peak'- the timing of the F0 peak being described with reference to the accented syllable). Thus, production data had been recorded using trained speakers practised in the production of patterns describable in terms of the hypothesized categories.

For perception experiments it was possible to produce all three positional variants of a single utterance by manipulating a naturally produced F0 peak, shifting it through a selected target utterance. It was found that the distinction between the three patterns is mainly due to a single variable: the temporal position of the peak relative to the vowel onset of the accented syllable.

Logically, the next step in the intonation project involved the intonation of German terminal utterances containing two pitch accents. An initial working hypothesis stated that the intonation of utterances with two (or more) peaks might be described as a sequence of simply concatenated single peaks. However, both initial production data, collected from trained speakers in various pilot tests, and results from various perception tests showed that the range of occurring categories exceeded the predictive capability of such a simple hypothesis, since:

a) Productions of such utterances show that sequences of peaks only rarely comprise 'classical' single-peaked configurations separated by a valley, individual accents often blending into a single 'wider' peak, or plateau (i.e. 'hat') configuration by some as yet poorly understood process of 'peak sandhi'.

1. Introduction

Besides the numerous theoretical difficulties involved in any analysis of intonation, the collection of naturally spoken data - with an eye to testing whatever theory has been developed - represents a considerable problem. Raw data recorded from an uncontrolled source outside the laboratory, eg. the radio, might satisfy the desire for 'naturalness', but are littered with unwanted artefacts and may be of inferior quality. On the other hand, controlled data recorded under laboratory conditions may boast a very high quality, but the stringency of the recording situation, heightened by the frequent need to obtain many repetitions of the 'same' intonation pattern(s), often proves too much for naive speakers. Much current research, then, is based on production data spoken by trained speakers, i.e. the researchers themselves or their associates. This fact clearly endangers the naturalness - in real terms - of the data, and casts doubt on the objectivity of any subsequent analysis.

The method described below aimed to elicit - from (potentially) any speaker and under good recording conditions - the desired intonation patterns spoken in a natural manner. Whilst the method in question was developed to test aspects of a particular intonation model, there can be no doubt that the method itself is relatively independent of any theoretical considerations governing the choice of the test material used. Hopefully, then, it should be of some interest to all intonation researchers.
b) Perception results have shown that a simple, linear relationship between peak height and perceptual prominence is insufficient as a model element; yet a certain prominence relationship between accents belonging to a single intonational unit might be relevant for a phonological categorisation of that unit.

c) Whereas in single-peaked utterances the temporal domain of the focus represented by the accent is more or less the whole utterance itself, two-peak utterances raise the problem of establishing the location of the border between the two accents' domains. Locating such a border may be of major importance for any phonological or phonetic description of the intonation contour; indeed, the two domains might even overlap. Pursuing the goal of extending a limited model based on single-peak utterances of a specific type to multi-peak configurations meant facing up the methodological problems sketched in the introduction as well as the obvious need for a more refined theoretical framework to match the complexity of the observed phenomena.

In principle, there are two complementary methods of establishing tentative intonational categories:

(i) perception experiments involving resynthesized material varied systematically in parameters relevant for the distinction of the categories, or

(ii) the analysis of production data.

The perception method can only be fruitful if we have at our disposal either some information about the relevant parameters involved or some fixed reference points in the potential domain of the contours represented by real productions. Otherwise we cannot guarantee that our resynthesized test material contains variations bearing on any intonational categories at all.

We decided to collect some production data which could serve to permit the formulation of hypotheses for further perception experiments. Initial attempts involved the recording of isolated sentences, read aloud from cards according to the standard laboratory paradigm. The aim of this was to obtain a wide range of intonational patterns, giving us some idea of the scope of variation possible. Thus, the speakers were only saddled with the bare minimum of instructions, to ensure that their utterances contained two accents. Surprisingly, perhaps, this unaccustomed articulatory freedom proved problematic for the speakers, who were more used to receiving extensive instructions in laboratory production situations. Hence, most utterances were produced with a single, safely stereotypical intonation pattern, instantly recognisable as such used for reading isolated sentences aloud, and consequently of little interest for our efforts to elicit natural, unconsciously controlled intonation patterns.

We therefore decided to adopt a method based on previous experiments using dialogue situations (see Gartenberg & Hertrich 1988) [1]. These experiments had been conducted using the following methods:

(1) Question and answer pairs: Subjects were quizzed 'live' by the experimenter about their age, name, nationality, and so forth according to a strictly prepared plan, involving the presentation of several, randomised repetitions of a limited set of utterances. They were faced either with questions serving as plain requests for the confirmation of basic information or assertions containing false information which they had to counter. Whichever was the case, the subjects were required to reply immediately, truthfully, avoiding the words "yes" and "no", and ensuring that their answers were complete. These restrictions served to control the linguistic content of the elicited responses and to motivate the speakers to express the pragmatics of the situation (eg. their agreement or disagreement) intonationally. By this method it proved possible to predict some intonation patterns produced by the subjects in terms of the pragmatic variable 'wrong/right assumption of the experimenter'.

(2) Spontaneous, live dialogues: These involved a situation in which subjects were asked basic personal questions data, produce some fixed official forms. These interviews were intended to evoke intonation patterns specifically connected with emphasis. The interviewer thus behaved in a forceful, contradictory manner, in an attempt to provoke emphatic responses. However, mainly due to the lack of situational and linguistic restrictions imposed upon the circumstances of the recording, the data elicited in this way proved too complex for analysis, though largely satisfying the desired naturalness criterion.

(3) Simulated speech recognition dialogues: Subjects replied to utterances prerecorded on tape containing preset pauses for the answers. The tape part, spoken by the experimenters in a machine-like manner, comprised not only questions but also some statements/interruptions such as "This information cannot be correct" or "Please speak more clearly", followed by a repetition of an earlier question. This procedure served to mimic the reaction of a fictional speech recognizer and to unsettle the subjects, the aim being to elicit emphatic utterances, hypothetically associated with 'late' peaks in the Kiel model. An auditory evaluation of these recordings suggested that the subjects certainly reacted to the 'machine', and that in a natural sounding manner. However, the data still proved too heterogeneous for either within- or cross-speaker comparisons.

In summary, we are faced with a kind of paradox: strict control of the subjects' responses (as in the first method above) detracts from the natural spontaneity of their productions; but insufficient control (methods 2 and 3) leads to problems of data heterogeneity. Some authors, eg. Cooper et al. (1985) [2] have tried to overcome the latter difficulty by using paper cards with the test sentences written on them. We have found, however, that these cards produced sentences (varying from noise on the recording to the provocation of 'reading' intonation), and that it is very difficult to keep the situation standardized if for each sentence a card has to be presented by the experimenter.

The task facing us, then, was to fix the text of the test utterances without destroying both the 'natural' dialogue atmosphere and the standardisation of the whole recording situation. In the following a 'new' method, allowing this, is described.
3. The new method

A dialogue is constructed comprising two distinct and alternating roles: Role one serves to set the desired situational framework for subsequent responses, and is prerecorded; role two consists of cued text, automatically presented briefly on a monitor screen at the appropriate moment in the dialogue, and is produced immediately by the speaker. Role one is electronically spliced onto tape for presentation to the 'live' conversation partner. Gaps tailored to the length of response and suitable to the dialogue structure are inserted between consecutive spliced utterances. Role two is recorded on a second tape for subsequent analysis. A microcomputer is used to synchronize the tape-recording of role one with the screen display of role two.

3.1. Technical design

TAPE 1 contains the prerecorded dialogue role on track one and control signals, consisting of sine waves, on track two. These signals, of different frequencies, serve respectively to mark the end of each utterance recorded on track one and the beginning/end of the dialogue.

The GENDET is a hardware generator and detector, used to generate and record the control signals onto tape as well as to detect them and relay them to the computer.

The monitor of the XT personal computer is used to present the line of text from the live dialogue role to the subject. The length of presentation time is crucial: It must be sufficiently long to ensure the retention of the text but also short enough to avoid its being 'read off' the screen in a stilted manner. When a signal is transferred by the GENDET, one line of a text file stored on disk is presented on the screen for a period which can be determined by the experimenter before the experiment is started; ca. 2 seconds is required for simple sentences. After presentation, the screen is cleared automatically so that no text is displayed while the speaker listens to the next utterance from their taped dialogue partner.

On TAPE 2 the sentences produced by the test speaker are recorded for further analysis.

4. An application of the method

Since the aim of our research was to evoke natural, but linguistically controlled, utterances containing two accents, a dialogue was constructed with the intention of evoking just such patterns in the answers to be spoken by the subjects. Linguistic control was achieved both by the semantic and intonal content of the prerecorded utterances, and by the temporal structure of the dialogue, building up as it proceeds, is extremely important, allowing back-references and contrasts, which not only serve to ensure the retention of the text but also pressure the subject into concentrating on the task in hand and really treating the dialogue material like a feasible and genuine exchange of information, and not as a purely phonetic exercise.

The dialogue used in the experiment was simple enough to enable the subjects to react spontaneously, but was - both grammatically and pragmatically - sufficiently varied so as to be unpredictable and provoke responses comprising a range of intonation patterns. For the subjects the situation resembled a telephone call with someone previously unknown to them and which consisted of requests for information mixed with subjective and sometimes contradictory comments.

The complete dialogue, which was in German, is given in the appendix. Recordings were made of 5 subjects, each of whom repeated the dialogue several times, at various intervals. Before they began the first time they were familiarised with the test situation by way of a short, preparatory dialogue similar to, but not taken from, the test dialogue. In almost every case, and despite the complexity of the recording situation, a few sentences sufficed. It was also noticeable that most subjects visibly enjoyed the task.

Since the major concern at this early stage of research was that the method succeed in producing natural, spontaneous utterances containing two accents, no strict analysis was immediately undertaken. Instead, a semi-formal auditory-phonetic evaluation was made, with the following results:

a) The experimenters checked whether the test utterances had been produced in a manner befitting the context. With very few exceptions this was indeed the case. Only on two or three occasions - in the first repetition - did a subject appear to 'make a mistake' (almost always immediately or subsequently acknowledged as such), either due to the pressure of the situation or to false anticipations. Thus the method does appear to work in principle, despite the unaccustomed mixture of audio and visual cues. Certainly the subjects' utterances did not sound 'mechanical' or 'read', even after several repetitions.

b) Three trained phoneticians judged how often the subjects' responses contained the two accents required. This varied from speaker to speaker, fluctuating between 90 and 100%.

c) In an informal test a group of listeners acquainted with the subjects confirmed that the subjects' responses had been "in character" as far as they could ascertain.
5. Conclusion

This method using a simulated dialogue boasts a combination of properties which cannot be said to apply to most standard procedures for collecting production data. These properties are:

(1) The recording situation is completely standardized and replicable.

(2) Artefacts from reading isolated sentences off cards are avoided.

(3) The same material can be spoken by many subjects without time-consuming preparations for each individual recording.

(4) Naive subjects require neither lengthy training nor extensive assimilation to the recording situation.

(5) The verbal context of the test sentences is uniformly fixed for all subjects.

(6) The subjects seem to enjoy the challenge and relative normality of the task.

Bibliography


Appendix

The test dialogue:

<table>
<thead>
<tr>
<th>Prerecorded role</th>
<th>Subject's response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welche Stifte hab' ich zur Auswahl?</td>
<td>Den blauen und den gelben.</td>
</tr>
<tr>
<td>Lieber wäre mir der grüne.</td>
<td>Der grüne ist vergeben.</td>
</tr>
<tr>
<td>Welche anderen Farben gibt es?</td>
<td>Lila und rosa.</td>
</tr>
<tr>
<td>Welcher malt am besten?</td>
<td>Malen tun alle.</td>
</tr>
<tr>
<td>Da bin ich nicht so sicher.</td>
<td>Die hier malen alle.</td>
</tr>
<tr>
<td>Auch der gelbe?</td>
<td>Der ist in Ordnung.</td>
</tr>
<tr>
<td>Was ist mit dem braunen?</td>
<td>Braun ist doch langweilig.</td>
</tr>
<tr>
<td>Und gehen tut er auch nicht?</td>
<td>Heut' morgen ging er noch.</td>
</tr>
<tr>
<td>Welche sind am teuersten?</td>
<td>Blau und gelb.</td>
</tr>
<tr>
<td>Der braue ist mir zu dick.</td>
<td>Der gelbe ist dünner.</td>
</tr>
<tr>
<td>Der ist auch nicht dünn genug.</td>
<td>Der grüne ist noch dünner.</td>
</tr>
<tr>
<td>Nein, er ist dicker.</td>
<td>Der ist doch dünner als alle</td>
</tr>
<tr>
<td>Der geht aber nicht</td>
<td>Der grüne geht.</td>
</tr>
<tr>
<td>Dann neh'n ich den.</td>
<td></td>
</tr>
</tbody>
</table>