Prosodic Imitation: 'Reliable' Listener Results

Duncan Markham
Department of Linguistics and Phonetics
Helgonabacken 12, S-22362 Lund, Sweden

ABSTRACT

Some of the results of an investigation of the imitation of auditory stimuli, and the identification of spontaneous and imitative speech are presented. Listeners were found to have difficulty identifying non-spontaneous stimuli, and judgements of imitative quality were not found to correlate with naïve spontaneous~non-spontaneous judgements.

INTRODUCTION

The experiment presented here is part of an investigation which seeks to determine whether it is possible to elicit perceptually and prosodically 'spontaneous' speech by asking informants to imitate auditory stimuli from a corpus of spontaneous speech. The results of the perceptual-assessment phase of the investigation were first presented in Markham (1993). In this paper I will present the results for those listeners in this phase who showed highest consistency in their judgements.

PROCEDURE

Two speakers of Swedish (one female, one male) heard complete or partial phrases excised from a recording of a one-hour quasi-spontaneous dialogue (the primary topic of conversation was prompted) involving the same speakers. The speakers were instructed to try to imitate what they heard. This procedure was repeated once, yielding two imitations per speaker for each of the stimuli. Further details can be found in Markham (1993).

Ten speakers of Swedish listened to the original and repetitions of ten (6 female, 4 male) of the sentences mentioned above. The investigation consisted of three stages, the first and third of which are discussed here. In Stage 1 listeners were told that they were to hear 100 utterances from two informants, some of which were spoken in a conversation, some in isolation, and were instructed to indicate for each utterance if they thought it might have occurred in conversation, or if it sounded as though it might not have been said spontaneously (*LAB*). 100 pseudo-randomized sentences (10 sentences x 5(original+4 repetitions) x 2) were presented via headphones approximately 1.4 seconds apart. The listeners indicated their judgement on a sheet of paper using the following score-field system for each sentence (score values for later assessment are shown in brackets): *LAB*(1), *PROBABLY LAB*(2), *PROBABLY CONVERSATION*(3), *CONVERSATION*(4).

In Stage 3 the listeners were informed that the first of every five 'identical' sentences was in fact the original, and were then asked to judge the quality of the imitations (*poor-OK-good*).

RESULTS

It was found that a number of listeners showed considerable unreliability in their judgements in Stage 1. This was observed in the discrepancies between scoring of the first and second instances of identical stimuli. Five of the ten informants appeared to differ from the rest in that they showed higher consistency in the scoring of identical stimuli. This was determined from grouping zero-field (same score) and one-field (eg. *LAB - PROBABLY LAB)* judgement differences together temporarily as one score category (Fig 1, Table 1). This seemed appropriate, as the nature and difficulty of the task (naturalness judgements) meant that some uncertainty in scoring was likely to occur.

Three-field differences, the greatest discrepancy possible, indicate that one member received the score *LAB*(1) whilst the second member was scored as *CONVERSATION*(4). Of the five more consistent informants (henceforth 'reliable'), two showed instances of three-field differences (Table 1), whilst all showed lower two-field difference scores than the

Table 1. *Scoring differences between members of 50 pairs of identical stimuli. The linguists in the group are starred, the non-Swedes are marked with two stars.*

INFORMANT	*1*	*3*	**7*	**8*	***9*
one-field	22	23	21	26	32
identical	24	22	26	18	16
TOTAL	***46***	***45***	***47***	***44***	***48***
three-field		1		2	
two-field	4	4	3	4	2
TOTAL	***4***	***5***	***3***	***6***	***2***

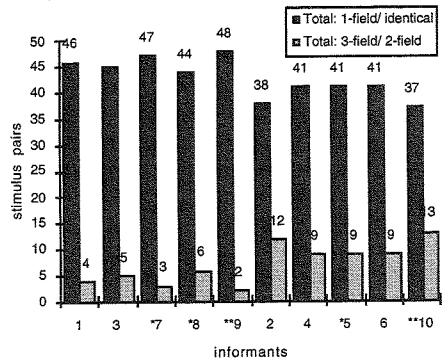


Figure 1. *Scoring differences for pairs of identical stimuli (reliable group to the left)*

informants I regard as less consistent.

Mean scores (where *LAB* = 1.0 and *CONVERSATION* = 4.0) for each stimulus ranged between 1.6 (1.5 for all ten informants) and 3.8 (3.5), whilst the standard deviation lay between 0.0 (0.48) and 1.52 (1.41). Differences in the mean score for identical members of a pair greater than or equal to 0.9 were observed in three (2) cases, and in nine (14) cases for differences below 0.9 down to 0.5. The remaining 38 (34) cases showed a difference in means less than 0.5.

The scores for 12 (75) of the 100 stimuli encompassed all fields, 22 (17) had three score fields (*LAB - PROBABLY LAB - PROBABLY CONVERSATION*), 33 (6) had three score fields (*PROBABLY LAB - PROBABLY CONVERSATION - CONVERSATION*), and 31 (2) had two score fields, whilst there were 2 (0) cases of unanimous scoring. This is markedly different from the results for all ten informants, shown in brackets. There is a drastic reduction in range of scores (field differences) (Fig 2), and a notable shift towards pro-conversation scores (Figs 3a, b — simple regression (Pearson's correlation coefficient, r). There is significant statistical correlation ($p \leq 0.0001$) between the reliable group and the remaining informants, although the real spread and lack of direct agreement in values is clear from visual examination of the figures.

The reliable group were not better than the remaining informants at being able to discriminate between real spontaneous stimuli and the copies, as can be seen in Table 2 (figures in brackets refer to all ten informants). The identification of the utterances as

Table 2. *Identification according to sentence of stimuli pairs for 10 sentences as CONVERSATION in Stage 1; 'identified' refers to pairs where members were scored as 4-4, 4-3, 3-3, and imitation stimuli were not; 'preferred spontaneous' indicates scores 4-4, 4-3, 3-3, where imitation stimuli also received similar, but not better scores. m=male, f=female. Stage 2 shows uniquely identified spontaneous stimuli for each speaker.*

SENTENCE	*1 f*	*2 f*	*3 m*	*4 f*	*5 m*	*6 m*	*7 f*	*8 f*	*9 f*	*10 m*
-preferred *spontaneous*				**2** [5]	**1** [3]	**1** [2]	**1** [1]	[4]		[1]
-identified	**1** [2]		**1**	[1]	**2** [3]	**2** [4]	**1** [1]	[1]	**2** [4]	**1** [3]
TOTAL	***1 [2]***	***0 [10]***	***1 [2]***	***2 [6]***	***3 [6]***	***3 [6]***	***2 [2]***	***0 [5]***	***2 [4]***	***1 [4]***
Stage 2	1 [3]	1 [2]	1 [2]	1 [3]	1 [3]	1 [5]	1 [3]	1 [2]	1 [2]	1 [5]
imitation chosen	[1]	[1]	**1** [2]	**1** [2]	**2** [4]	**1** [3]	[2]		[1]	**3** [4]
none chosen	**4** [7]	**5** [9]	**3** [6]	**2** [2]		**1** [1]	**3** [6]	**5** [5]	**3** [5]	**1** [2]

CONVERSATION for the reliable group is shown in Table 2 and Fig 4. No clear improvement in the number of correctly identified stimuli is apparent for the more consistent informants, indeed the identification of spontaneous stimuli is actually worse.

In Stage 3 imitations were scored for imitative quality. 6 (28) of 40 sentences received scores in all three fields (*poor*(1)-*OK*(2)-*good*(3)). 12 (6) sentences were scored as *OK* or *good*, 18 (6) as *poor* or *OK*. 4 (0) sentences received scores in one field only (always *OK*). Interestingly, this is a definite increase in agreement across informants, despite the fact that agreement in identification did not improve for the reliable group in Stage 1 (cf Table 2).

Only one of the reliable group showed a significant correlation (Spearman Rank Correlation) between scores given in Stage 1 and quality judgements given in Stage 3, and even then the significance level was only at $p \leq 0.05$. Comparison of judgements across informants for both Stages 1 and 3 revealed some similarities. In Stage 1, two discrete pairs of informants in the reliable group were identified as giving similar scores with significance at $p \leq 0.0001$. However, a cross-pair comparison did not show any correlation in scores. In Stage 3, three informants gave significantly correlated quality judgements at $p \leq 0.0002$. All except one pair comparison yielded relatively low p-values, indicating a higher degree of agreement regarding the quality of the imitations, than regarding the identification of the imitations as such, as the p-values obtained for Stage 1 were generally higher.

In Markham (1993:97) I suggested testing the quality of the imitations by splicing them into the original dialogue. In an informal test, two sections of

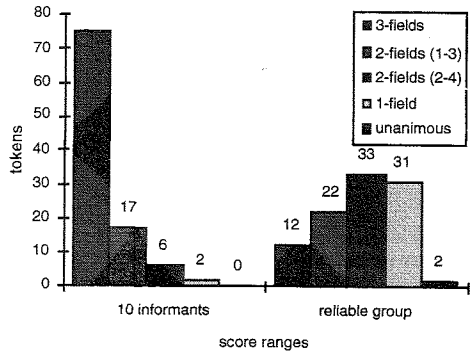
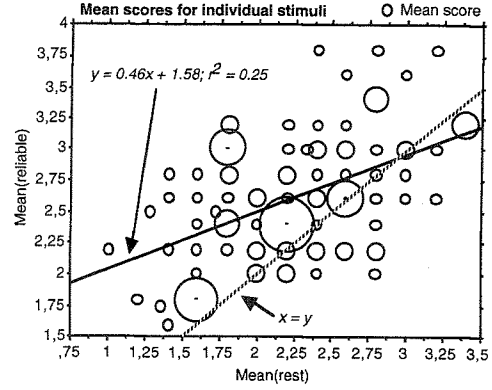


Figure 2. *Score ranges (variation in scoring) for individual tokens across all informants.*



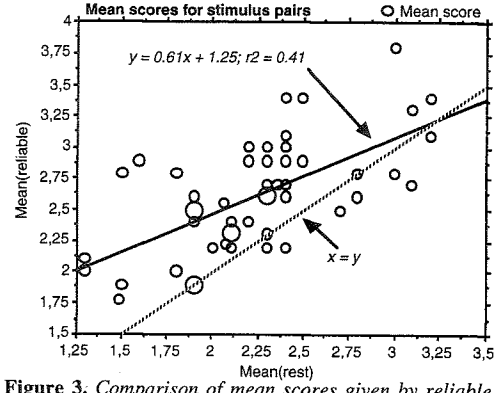


Figure 3. *Comparison of mean scores given by reliable and other informants for individual and pairs of stimuli. Larger points show point coincidence at given positions.*

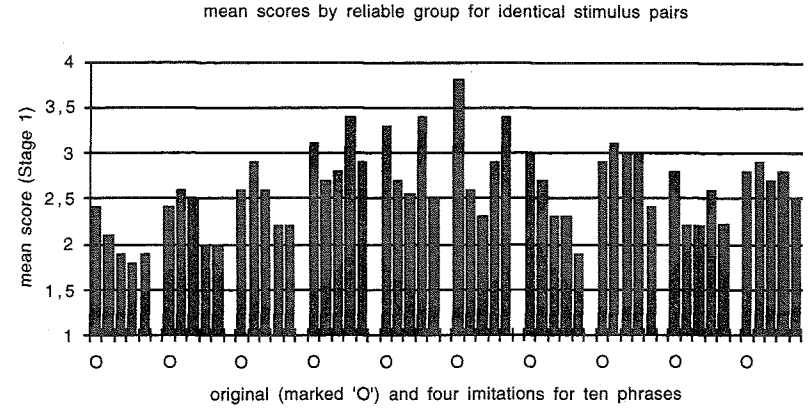


Figure 4. *Mean scores given by the reliable group for identical stimulus pairs (the original and four imitations of ten phrases) (Stage 1). 1=LAB, 4=CONVERSATION.*

dialogue were presented to a group of listeners. Three versions of each section were presented, each containing either the original, the first, or the second imitation by the female speaker. The listeners, many of whom had participated in the investigation described above, could not discriminate between the sections containing the original and those containing imitations. This does not necessarily mean that the speaker had achieved a good imitation, but at the very least that the non-spontaneous utterances (imitations) were in their own right acceptable as 'spontaneous'. The fact that they still fitted into the original dialogue indicate that the speaker achieved an acceptable imitation of the initial and final pitch level to give the impression of 'pitch concord' — the approximate matching of pitch between two speakers at a dialogue turn-juncture (cf Botinis (1992:45f). Such productional issues are discussed in Markham (1994 forthc).

SUMMARY

The results presented in this paper provide strong evidence that the perceptual distinction between spontaneous and imitative utterances is by no means clear. Utterances judged by the informants as being only 'OK' or 'poor' imitations (Stage 3) could only seldom be identified as clearly non-spontaneous utterances (Stage 1), whilst conversely, spontaneous utterances were not always identified as such (cf Fig. 4), with some imitations actually receiving higher scores. No correlation was found between the scores in Stages 1 and 3. Some similar behaviour was found amongst the informants, but no uniform cross-listener strategy could be observed. Despite the fact that this paper focused specifically on the reliable listeners amongst the ten original informants, no clearer tendencies could be identified. The reliable listeners appear merely to be better at making consistent judgements, without the judgements themselves being more informative *per se*. This does, however, provide a firmer basis for an investigation of the productional correlates of the perceptual impressions (Markham 1994 forthc).

REFERENCES

Botinis, A. (1992): 'Accentual distribution in Greek discourse'. *Travaux de l'Institut de Phonétique d'Aix,* 14: 13-52.

Markham, D.J. (1993): 'Prosodic imitation: perceptual results'. In J.S. Pettersson (ed.), *Seventh Swedish Phonetics Conference (Fonetik-93).* Reports from Uppsala University, Dept of Linguistics (RUUL#23), 94-97.

Markham, D.J. (1994 forthc): 'Prosodic imitation: Productional results.'