THE EFFECT OF REMOVING SEMANTIC INFORMATION UPON THE IMPACT OF VOICE IMITATION

Kirk P H Sullivan [1, 2], Elisabeth Zetterholm [3], Jan van Doorn [4],
James Green [5], Frank Kügler [6] & Erik Eriksson [1, 2]

[1] Department of Philosophy and Linguistics, Umeå University, Sweden
[2] Department of Computing Science, Umeå University, Sweden
[3] Department of Linguistics and Phonetics, Lund University, Sweden
[4] Department of Clinical Sciences, Umeå University, Sweden
[5] Department of Psychology, Otago University, New Zealand
[6] Department of Linguistics, University of Potsdam, Germany

ABSTRACT: Previous research has shown both that listeners' ability to detect high quality voice imitation results in judicially worrying misidentification rates (Schlichting & Sullivan, 1997) and that the semantic expectation of the listeners as to the content of a message impacts upon the acceptance of a voice imitation (Zetterholm et al., this volume). This paper considers whether the results of the Zetterholm et al. paper were due to a difference in imitation quality that was not detected in the phonetic analyses of the imitations rather than the semantic expectations of the listeners. The earlier study was replicated, using different listener groups with various linguistic backgrounds and levels of familiarity with the voice that was imitated. The results reveal a complex interaction of language background and familiarity with the target voice, that supports the previous finding, yet suggests that the impact of semantic expectation may be lower than posited in Zetterholm et al.

INTRODUCTION

High quality speaker imitation has been shown to lead to speaker misidentification rates that are of judicial concern (Schlichting & Sullivan, 1997). A recent study by Zetterholm et al. (2002) found that listener expectation relating to the topic of the imitated passage has a strong impact upon the acceptance or rejection of an imitation. The results of that study showed that listeners were more likely to accept a voice imitation as the actual voice if the topic of the speech passage was consistent with listener expectations. In that study two imitations of a single famous voice by the same imitator were used to investigate the importance of semantic expectation upon the acceptance of an imitated voice. The imitations were semantically extremely different but were judged by three well-trained phoneticians to be phonetically equivalent. Further, the imitator was judged to have successfully imitated the primary distinctive characteristics of the voice being imitated.

These auditory judgements of equivalence were, however, subjective and it is possible that the phonetic analysis could have missed important differences in the quality of the imitations that were, in fact, the cause of the findings in the earlier research, rather than the interpretation given in terms of listener expectations. In order to investigate whether any important differences between the two imitations had been missed in the earlier phonetic analyses and had led to the results reported in Zetterholm et al. (2002), the earlier study was replicated, using different listener groups with various linguistic backgrounds and levels of familiarity with the voice that was imitated.

In the first study, the listeners were all Swedish speaking adults. In the replicated study five listener groups have been chosen: adult Swedish speakers (similar to the original study), teenage Swedish speakers, Danish speakers, German speakers, and New Zealand English speakers. The adult Swedish speaking group was chosen to allow comparison with the first study, while the teenage Swedish group was likely to have been less familiar with the voice being imitated. The Danish speaking group would be able to understand the content of the training passages (the Scandinavian languages, Swedish, Norwegian and Danish are considered to be mutually intelligible (see e.g. Elert, 1981)), yet would be less aware of the phonetic characteristics of the imitated voice than either of the Swedish groups. The remaining two groups would have no knowledge of Swedish (or any other Scandinavian language). Further all three of the non-Swedish groups' perception of the voices would differ from the native Swedes due to their first language phonology and in the case of the non-Swedish understanding groups, the New Zealand and German groups, the broader characteristic traits of the voice would be relatively more important. The purpose of the present study is to use the responses by these groups of listeners to investigate in more detail the original hypothesis and finding, that the topic of the imitated passage has a strong impact upon the acceptance or rejection of a voice imitation.

METHOD

The method is a replication of the Zetterholm et al. (2002) experiment, except that a larger number of listener groups with different linguistic backgrounds was used. The replicated sections of the original method are summarised below, but the characteristics of the listener groups are described in detail.

THE VOICES

The voices consisted of a set of seven recordings of the same text of a political speech, and one recording of a non-political topic (how to bake a cake). The political recordings were an original by Carl Bildt (PS-Bildt), a professional imitation of the voice of Carl Bildt (PS-AMBildt), the natural voice of the professional imitator (PS-AM) and four other male voices, referred to hereafter as 'foils'. The non-political recording was a free voice imitation of Carl Bildt explaining how to bake a cake (H2BC-AMBildt). Two of the recordings (PS-AMBildt and H2BC-AMBildt) were used as familiarisation passages, and the remaining six were used as the basis for the voice line-up. Detailed information about the voices used in the experiment, along with the findings of the acoustic and auditory analysis of the recordings, can be found in Zetterholm et al. (2002).

PROCEDURE

Experimental participants: the listener groups

The listener groups were all randomly selected and no listener reported any hearing damage. There were five different listener groups. Each group was divided into two sub-groups, one sub-group undertook Experiment 1 and one sub-group undertook Experiment 2.

The listener groups were an adult native speaking Swedish group from Southern Sweden, a teenage native speaking Swedish group from Southern Sweden, a native Danish speaking group from Copenhagen, a native German speaking group from Potsdam and a native New Zealand English speaking group from Dunedin. After the perception experiments, all the participants reported whether they were familiar with the voice of Carl Bildt (FCB) or not, and if they were familiar with any of the other voices used in the experiment or not. The descriptive statistics of the listener groups is reported in Table 1.

Table 1: The descriptive statistics of the listener groups.

Group	Sub-group	No.	No. Male	No. Female	Mean Age	SD Age	% FCB
SE Adult	PS-AMBildt	27	10	17	43.96	13.12	96
	H2BC-AMBildt	27	10	17	44.19	9.60	96
SE Teen	PS-AMBildt	25	6	19	15.24	1.21	72
	H2BC-AMBildt	14	5	9	16.07	1.27	43
Danish	PS-AMBildt	36	2	34	25.28	5.74	3
	H2BC-AMBildt	29	12	17	25.72	7.26	30
German	PS-AMBildt	11	0	11	23.82	3.52	0
	H2BC-AMBildt	12	3	9	25.00	3.16	0
New Zealand	PS-AMBildt	21	8	13	18.60	0.58	0
	H2BC-AMBildt	14	7	10	19.23	2.08	0

Familiarisation and line-up speech material

Two experiments were constructed, each consisting of a familiarisation voice and six test voices. The experiments differed only in the familiarisation passage. Experiment 1 used the political passage imitation (PS-AMBildt), while Experiment 2 used the "how to bake a cake" passage imitation (H2BC-AMBildt). In both experiments the line-up was constructed from the six recordings of the political passage that were not used for the familiarisation task. Three separate segments that contained political content were spliced out from each recording. Each speech segment was repeated three times in the line-up, giving a total of 54 speech stimuli in the line-up (3 speech samples x 3 repetitions x 6 speakers). The line-up voices thus contained PS-Bildt, PS-AM and foils as the test voices.

Recognition tasks

The two experiments were conducted separately. Both sub-groups were first familiarised with the voice of the target speaker they were to identify, and were told that they would be asked to recognise it later. Then they listened to a CD containing the 54 speech stimuli, presented in a randomised order. The listeners were instructed to respond 'Yes' on a response sheet whenever they recognised the voice from the familiarisation recording, and 'No' when they did not.

DATA ANALYSIS

The Yes-No data were analysed using the same techniques used by Zetterholm et al. (2002). The data were categorised either as hits ('Yes' response to target voice stimulus), miss ('No' response to target voice stimulus), false alarm ('Yes' response to non-target voice stimulus) and correct rejection ('No' response to non-target stimulus) (Green & Swets, 1966).

By taking the number of hits and false alarms, together with the total number of target stimuli and non-target stimuli presented, it is possible to calculate the listeners' discrimination sensitivity, d', as follows, d' = z(H) – z(FA), where z() represents the transformation of a proportion to a z-score, H represents the proportion of target trials where the listeners scored a 'hit' and FA represents the proportion of non-target trials where the listeners scored a 'false alarm'.

A measure of response bias, c, (c = –0.5[z(H) + z(FA)]) was also calculated to check whether listeners had a tendency to answer 'yes' in preference to answer 'no', or vice-versa. A positive c value indicates a preference to answer 'no'; here the false alarm rate is lower than the miss rate. A negative c value indicates a preference to answer 'yes'; here the false alarm rate is greater than the miss rate.

RESULTS

Following Zetterholm et al. (2002), hits and false alarms were counted for the participants and then pooled. The responses were counted first with the imitator's (AM's) voice as the target voice, (that is identification of the natural voice of the imitator being scored as a hit), and then with the voice of Carl Bildt, voice imitated, as the target voice. The first set of scores shows how well, or badly, the imitator succeeded in convincing the listeners that he was someone else by imitating the voice of Carl Bildt. The second set of scores show how well he convinced the listeners that he was Carl Bildt. Table 2 shows the distribution of responses for the listener groups who heard PS-AMBildt as their training passage, Table 3 shows the distribution of responses for the listener groups who heard H2BC-AMBildt as their training passages and Table 4 summarises the d' and c-scores for all the listener groups.

DISCUSSION

The impact of semantics was less dramatic for the adult Swedish group than the results reported in Zetterholm et al. (2002). However, the impact of the different training passages was clear in the d' values when CB was scored as the target voice. For the PS-AMBildt familiarisation passage, d' is 2.272, which is higher than the d' value (1.1100) for the H2BC-AMBildt familiarisation passage. The response bias values of 0.439 and 0.535, are similar and in both cases the false alarm rate is lower than the miss-rate (see Table 4). The similarity of the response biases strengthens the d' result, that

Table 2: Distribution of responses for Experiment 1 listeners, who heard PS-AMBildt. These are scored for AM as the target to the left and CB to the right. The hit rate (H) and false alarm rate (FA) are also shown.

Listener Group	Stimulus class	AM target voice Response			CB target voice Response		
		'Yes'	'No'	Total	'Yes'	'No'	Total
Swedish Adult	Target	12 (H = 0.05)	231	243	213 (H = 0.76)	30	243
(n = 27)	Non-target	242 (FA = 0.19)	973	1215	70 (FA = 0.06)	1145	1215
Swedish Teens	Target	12 (H = 0.05)	213	225	134 (H = 0.60)	91	225
(n=25)	Non-target	250 (FA = 0.22)	875	1125	128 (FA = 0.11)	997	1125
Danish	Target	26 (H = 0.08)	298	324	130 (H = 0.40)	194	324
(n=36)	Non-target	366 (FA = 0.23)	1254	1620	262 (FA = 0.16)	1358	1620
German	Target	45 (H = 0.45)	54	99	55 (H = 0.56)	44	99
(n=11)	Non-target	237 (FA = 0.48)	258	495	227 (FA = 0.46)	268	495
New Zealand	Target	46 (H = 0.24)	143	189	117 (H = 0.62)	72	189
(n=21)	Non-target	459 (FA = 0.49)	486	945	388 (FA = 0.41)	557	945

Table 3: Distribution of responses for Experiment 2 listeners, who heard H2BC-AMBildt. These are scored for AM as the target to the left and CB to the right. The hit rate (H) and false alarm rate (FA) are also shown.

Listener Group	Stimulus class	AM target voice Response			CB target voice Response		
		'Yes'	'No'	Total	'Yes'	'No'	Total
Swedish Adult	Target	11 (H = 0.05)	231	243	123 (H = 0.51)	120	243
(n = 27)	Non-target	283 (FA =0.22)	932	1215	16 (FA = 0.13)	1199	1215
Swedish Teens	Target	8 (H = 0.05)	118	126	67 (H = 0.53)	59	126
(n=25)	Non-target	163 (FA = 0.26)	467	630	104 (FA = 0.17)	526	630
Danish	Target	39 (H = 0.34)	222	261	88 (H = 0.15)	173	261
(n=36)	Non-target	486 (FA = 0.37)	819	1305	422 (FA = 0.32)	883	1305
German	Target	34 (H = 0.32)	74	108	34 (H = 0.32)	74	108
(n=11)	Non-target	152 (FA = 0.28)	388	540	152 (FA = 0.28)	388	540
New Zealand	Target	45 (H = 0.36)	81	126	80 (H = 0.63)	46	126
(n=21)	Non-target	305 (FA = 0.30)	325	630	169 (FA = 0.25)	461	630

the listeners were more ready to accept CB as the voice they had been asked to remember having heard PS-AMBildt as the familiarisation text. A similar, but less marked result was found for the teenage listeners and a particularly small difference was found for the Danish listener group. In none of these three groups, however, was a major impact of the training passage found in detection of the voice of AM, when scored as the target voice. A clear distinction was reported in Zetterholm et al. (2002); PS-AMBildt lead to a larger negative d' than H2BC-AMBildt. In the case of the three Scandinavian groups in this study H2BC-AMBildt lead to larger negative d' values when AM was scored as the target voice.

The difference in responses between the adult and teenage groups of Swedish listeners most likely can be attributed to the difference in familiarity with the target voice CB. CB lost the 1994 General election and thereafter became gradually less central to Swedish politics. Many of the teenage listeners were familiar with the voice of CB, yet would have had far less exposure to his voice than the adult group. Thus, the semantics of the training passage had primed recognition of CB, yet the specific features of

Table 4: Summary of the d’ and c values for Experiments 1 and 2.

Listener Group	Training Passage	Target Voice	d’	c
Swedish Adult	PS-AMBildt	AM	-0.806	2.272
	H2BC-AMBildt		-0.933	1.226
	PS-AMBildt	CB	2.272	0.439
	H2BC-AMBildt		1.100	0.535
Swedish Teens	PS-AMBildt	AM	-0.845	1.189
	H2BC-AMBildt		-0.879	1.054
	PS-AMBildt	CB	1.449	0.482
	H2BC-AMBildt		1.087	0.447
Danish	PS-AMBildt	AM	-0.651	1.078
	H2BC-AMBildt		-0.713	0.381
	PS-AMBildt	CB	0.737	0.619
	H2BC-AMBildt		0.682	0.439
German	PS-AMBildt	AM	-0.061	0.084
	H2BC-AMBildt		0.962	0.530
	PS-AMBildt	CB	0.244	-0.018
	H2BC-AMBildt		0.962	0.530
New Zealand	PS-AMBildt	AM	-0.660	0.366
	H2BC-AMBildt		-0.322	0.529
	PS-AMBildt	CB	0.529	-0.039
	H2BC-AMBildt		0.205	-0.080

the voice were less clear for the teenagers. The importance of the interaction of semantic expectation and degree of familiarity with the voice for accepting a voice imitation, is further suggested by the teenagers’ d’ values when AM was scored as the target voice. Here the impact of the semantics of the training passage was minimal; the d’ values lie between those for the adult groups. The gap between the d’ values for the teenager group was less than that for the adult group. Yet, in both cases successful recognition of AM was higher after hearing PS-AMBildt as the training voice. This does not reduplicate the finding of Zetterholm et al. (2002), where there was a clear preference against selecting AM after the PS-AMBildt training passage. A possible explanation for the differences between the two adult groups is that one group came from the north of Sweden and the other from the South of Sweden. The group from the south of Sweden spoke a variety of Swedish with more common features to that spoken by CB and this perhaps made the group better able to detect features that were inconsistent with the voice of CB and other feature that AM had altered in his imitation of CB that led them not to select AM’s voice. Another possible explanation is that the Southern Swedish group are older than the Northern Sweden group and thus more likely to be aware of the political era when CB was politically most active.

Interestingly, the Danish group’s responses resulted in d’ values that are nearer zero than those of the Swedish speaking groups, yet they maintain the pattern found in the Swedish speaking groups. Further this group of listeners’ d’ values show a greater impact of the semantic content of the training passage for the acceptance of CB as the target voice than was found in the teenage data. This is of importance as 3% of the listeners hearing the PS-AMBildt training text and 30% of the listeners hearing the H2BC-AMBildt training text reported that they were familiar with Carl Bildt’s voice. This combined with the reported results suggests that an impact of semantic communality between training passage and the semantics of the test voices needs to be factored out of the Swedish adult data.

In the listener groups that understood no Swedish, or other Scandinavian language, and whose responses, therefore, cannot be impacted upon by the semantics of the training passage, there are contradictory outcomes. The New Zealand group’s d’ values are nearer zero than the Scandinavian groups' and have a higher d' value when CB is scored as the target voice for the PS-AMBildt training passage (0.529 cf. 0.205), coupled with near zero negative c values. The German group's results are the opposite, with a higher d' after having listened to the H2BC-AMBildt training passage (0.244 cf. 0.962) coupled a c value of 0.530, ie a preference to answer 'no', for the H2BC-AMBildt training passage and near zero (-0.018) for the PS-AMBildt training data. The difference between the d' values for the two German subgroups when scoring AM as the target voice is also marked (-0.061 cf. 0.962) and contradicts that found in the New Zealand data (-0.660 cf. –0.322). Research by Schiller et al. (1997) found that removing semantic information resulted in listeners from different backgrounds performing similarly in speaker identification tasks. Although both groups made a clear distinction between AM and CB, the impact of the quality of the imitation, when the semantic content of the passages was not available to the listeners, did not result in the same preferences by the German and New Zealand listener groups. It is, thus, difficult to judge whether there was an undetected difference in recording quality in auditory and acoustics analyses reported in Zetterholm et al. (2002).

CONCLUSIONS

The results here provide an initial insight into the complex interaction of semantic expectation, familiarity with a voice, dialect, and first language. The results from the Danish group indicate an impact of semantic similarity between the training and test passage and the non-Scandinavian groups do not reveal a stable pattern that can be attributed to imitation quality. However, although the results presented here reveal a complex interaction of language background and familiarity with the target voice, they lend support to the proposition made by Zetterholm et al. (2002) that semantic expectation is an integral factor in the acceptance of an imitated voice. The results here, however, suggest that the degree of impact may be lower that found in the earlier reported Zetterholm et al. study.

A follow up study is currently investigating in more detail the importance of dialect and age within Swedish listeners on the impact of semantics on the acceptance of an imitated voice and more non-Scandinavian data is being collected. Further, it is probable that individual differences in listeners d' and c values are large and, thus, the results presented in this and the Zetterholm et al. (2002) papers should be confirmed by calculating individual rather than group d' and c values.

REFERENCES

Elert, C.-C. (Ed.) (1981) *Internordisk språkförståelse (Mutual Language Understanding in the Nordic Countries)*, Acta Universitatis Umensis, 33, Umeå University, Umeå Sweden.

Green, D. M. & Sweets, J. A. (1966) *Signal Detection Theory and Psychophysics*, New York: Wiley.

Schlichting, F. & Sullivan, K. P. H. (1997) *The imitated voice – a problem for voice line-ups?* Forensic Linguistics, 4(1): 148–165.

Schiller, N. O., Köster, O. & Duckworth, M. (1997) *The effect of removing linguistic information upon identifying speakers of a foreign language*, Forensic Linguistics, 4(1): 1–17.

Wretling, P., Sullivan, K. P. H. & Schlichting, F. (1999) *Does repeated exposure to a target voice reduce the impact of a similar voice?* Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, 1385–1388.

Zetterholm, E., Sullivan, K. P. H & van Doorn, J. (2002) *The impact of semantic expectation on the acceptance of a voice imitation*, this volume.