THE EFFECT OF VOICE SIMILARITY ON STREAM SEGREGATION

Angelika Braun & Helen Hahn

University of Trier, Germany
brauna@uni-trier.de; Helen-hahn@web.de

ABSTRACT

This contribution addresses the influence of voice similarity on a shadowing experiment. Two groups of participants (implicit-knowledge and novel-voice) had to shadow a target voice under two conditions, i.e. (1) when the target voice and the distracter voice were acoustically similar and (2) when they were dissimilar.

The error rates did not differ significantly between the two groups, but for both groups, the number of errors was significantly larger in the dissimilarity condition. This demonstrates that voice similarity is a factor to be considered in shadowing experiments.

Keywords: speech perception, stream segregation

1. INTRODUCTION

In order to gain insight into the speech perception mechanism the differences in processing between familiar and unfamiliar voices have been studied. Generally, processing advantages for familiar voices were found [3, 7, 8]. Birkett, et al. [1] were able to specify a certain area within the auditory cortex which was activated by a familiar voice.

On a different strand, the process of singling out one speaker against a background of noise has long been a subject of study; for a review cf. [4]. The worst kind of noise which can interfere with speech recognition is that of a second speaker. This phenomenon is well-known as cocktail-party-effect [2].

A recent study by Newman and Evers [7] combines the two aspects, adding the issue of implicit vs. explicit talker familiarity. While they found a significant advantage in performance (i.e., fewer errors) for the explicit-knowledge group in a shadowing experiment, no difference could be established between their implicit knowledge and novel voice groups.

The present research was prompted by the Newman and Evers study. It addresses still a different aspect of the shadowing task: Given the previous finding that physical dissimilarity between sound streams facilitates stream separation in non-speech signals [5], we hypothesized that the same may be true if a second voice is used in order to create "noise" in a shadowing task. If this proved to be the case, the (acoustic) difference between the distracter voice and the target voice might affect listener performance and thus the results of the segregation task. We decided to compare an implicit-knowledge group and a novel-voice group only because it is very clear that explicit knowledge facilitates stream segregation. The research questions asked here were:

Is there a difference in performance between the implicit-knowledge and the novel-voice groups?

Does it make a difference whether the target voice and the masking voice are similar with respect to fundamental frequency?

2. MATERIALS AND METHODS

2.1. Target speaker

In order to match the methodological approach chosen by Newman and Evers [7], a target speaker had to be found who was known to the implicit-knowledge (henceforth IK) group of listeners. An instructor from the phonetics department at the University of Trier who had been teaching one class per semester for the past 18 months served as speaker. He was 30 years old at the time of the recording and did not smoke or suffer from any voice, speech or language disorder. His pronunciation was close to Standard German with a slight regional accent.

The target speaker's recordings were analysed with respect to median F0, F0 standard deviation (based on 40 sec of read speech) as well as jitter and shimmer (based on a sustained vowel /a/ without on- and offglide) using the Praat software package. Results show a median F0 of 150 Hz with a standard deviation of 25 Hz, the coefficient of variation thus amounting to 0.16. While the median F0 is slightly higher than average for male speakers of German, the coefficient of variation indicates normal voice modulation. Jitter (rap) was found to be 0.098; shimmer (apq11) 1.60%. Both these values are distinctly non-pathological.

Speaking tempo was not measured, but the auditory impression did not suggest any anomalies.

2.2. Background speakers

Two background speakers were recorded who both resembled the target speaker in that they were male non-smoking native speakers of German, and did not exhibit any voice, speech or language disorder.

One background speaker (henceforth B1) differed from the target speaker with respect to age and median F0. He was 53 years old at the time of the recording, and his median F0 was measured at 95 Hz. The coefficient of variation (0.18) was slightly higher than that of the target speaker; jitter and shimmer were non-pathological.

The second background speaker was more similar to the target speaker in age and pitch. He was 23 years old at the time of the recording. His median voice fundamental frequency was measured at 131 Hz with a standard deviation of 20 Hz, the coefficient of variation being 0.15. Jitter and shimmer were non-pathological. Table 1 summarizes the speaker details.

Table 1: Speaker characteristics.

Characteristics	Target speaker	Distracter voice B1	Distracter voice B2
Sex	Male	Male	male
Age	30	53	23
F0 (median)	150.2 Hz	95.5 Hz	130.9 Hz
SD (Hz)	24.7	17.4	20.0 Hz
Variation	0.16	0.18	0.15
Shimmer (apq11)	1.600 %	1.581 %	3.662 %
Jitter (rap)	0.098 %	0.160 %	0.202 %

2.3. Stimuli

The speech material consisted of four short stories of about 3 min. duration each, authored by a popular German comedian [6]. The target speaker recorded two stories as well as the phrase "*Folgen Sie meiner Stimme*" ('follow my voice') and a sustained vowel /a/. The background speakers each recorded two different stories from the same book as well as sustained /a/ vowels. All recordings were made using a Thinkpad R60 laptop computer and a headset Sennheiser PC161. The recordings were digitized at 16bit 44.1kHz mono.

Two test tapes were created by binaurally blending the voice of the target speaker with that of one background speaker at a time, with the level of the target speaker's recordings exceeding that of each distracter by 5dB on average. The blended passage was preceded by three repetitions of the target voice saying "follow my voice".

2.4. Participants

A total of 25 participants, all of them university students, took part in the experiment. Thirteen of them knew the target speaker because they had attended his classes (implicit-knowledge group). The remaining 12 participants were tested at a different university and definitely did not know the target speaker.

Participants were asked to shadow the voice which instructed them to do so at the beginning of the recording. They were given a chance to get used to the task by listening to a test tape lasting 52 seconds which contained two female voices. The stimuli were presented over circumaural headphones in order to prevent distraction by the participant's own voice. The shadowed speech was recorded onto the hard disk of an IBM Thinkpad R60, using a headset Sennheiser PC 161.

After having listened to the training tape, participants were presented with the test tapes. There was a brief break between the two shadowing sessions, the whole experiment lasting about 20 min. In order to test for learning effects, half of the participants listened to the similar-voice condition (target speaker + B2) first, the other half started out with the dissimilar-voice condition (target speaker + B1). Upon completion of the experiment, each participant had to fill out a questionnaire. They were asked whether they knew one of the stories or one of the speakers. They were also asked whether they had made an effort to identify the speaker.

2.5. Scoring

The second minute of about three minutes of shadowing was analyzed for each participant. This selection was made in order to avoid the initial phase of getting used to the task and the final phase which could have been affected by fatigue. As in [7], three types of error were distinguished: wrong words (henceforth WW), missed words (MW), and insertions (I).

3. RESULTS

None of the participants knew any of the stories. (This would have been a reason for exclusion from the analysis.) Twelve out of 13 members of the implicit-knowledge group reported that it had not taken them long to be fairly sure of the target speaker's identity. Neither group reported to have been distracted by attempting to identify the speaker.

3.1. Implicit-knowledge group

Table 2 and Figure 1 summarize the results. The dissimilar voice scenario is represented by A1, the similar voice scenario by A2. Two out of three error types are more frequent in the similar-voice condition, the exception being the number of insertions. The bulk of errors – as in [7] – are missed words (MW); therefore statistical analysis focused on this type of error.

3.2. Novel-voice group

The results for the novel-voice group are contained in Table 3 and Figure 2. The absolute numbers of errors as well as the distribution over error types closely resemble those of the implicit-knowledge group.

Table 2: Descriptive statistics for the *implicit-knowledge* (IK) group (N = 13).

	Variable	*Mean*		*SD*
A1	WW	5.23		3.65
	MW	**15.38**		**16.89**
	I	4.85		3.21
A2	WW	6.23		4.46
	MW	**34.38**		**18.70**
	I	4.00		2.52
	Variable	*Min*	*Max*	*Range*
A1	WW	0	12	12
	MW	**1**	**47**	**46**
	I	0	11	11
A2	WW	2	19	17
	MW	**9**	**70**	**61**
	I	1	10	9

Figure 1: Error patterns of the implicit-knowledge group.

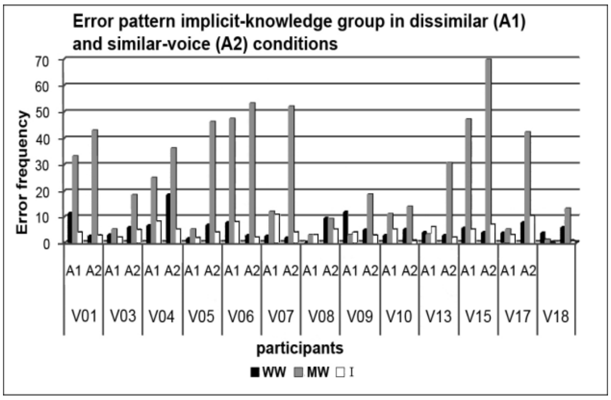
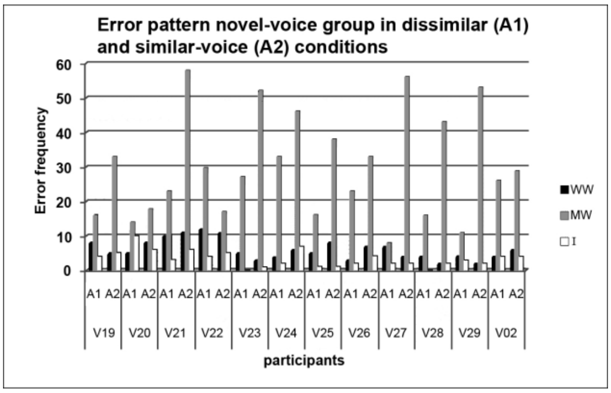


Table 3: Descriptive statistics for the *novel-voice* (NV) group (N = 12).

	Variable	*Mean*		*SD*
A1	WW	5.92		2.78
	MW	**20.25**		**7.86**
	I	2.92		2.64
A2	WW	6.08		3.09
	MW	**39.67**		**14.1**
	I	3.75		2.09
	Variable	*Min*	*Max*	*Range*
A1	WW	3	12	9
	MW	**8**	**33**	**25**
	I	0	0	10
A2	WW	2	11	9
	MW	**17**	**58**	**41**
	I	1	7	6

Figure 2: Error patterns of the novel-voice group.



3.3. Statistical analysis

The mean values of the missed word errors were subjected to statistical analysis. Two-tailed *t*-tests for independent samples were carried out where the data were normally distributed; when this was not the case (implicit-knowledge group; dissimilar voice condition) the non-parametric Mann-Whitney U-test was applied.

The absolute number of MW errors in the novel-voice group exceeded that of the implicit-knowledge group. However, neither in the similar-voice condition nor in the dissimilar-voice condition was there a significant difference between the two groups with respect to the number of missed words. (Similar-voice condition: 2-tailed *t*-test for independent samples with p = 0.943 and df = 24; dissimilar voice condition: Mann-Whitney U-test with p = 0.0985.) This result confirms the findings by Newman and Evers [7] in that there is no significant difference between the two groups and, at the same time, is at variance with them in that those authors found more errors in the implicit-knowledge group. This behavior can possibly be explained by the fact that the majority of our participants who knew the target speaker identified him correctly at an early stage of the experiment. Those who were certain of his identity could thus have benefited from this knowledge.

On the other hand, both groups made vastly more errors in the similar-voice condition than when two dissimilar voices were blended. For the implicit-knowledge group, p = 0.0046 (Mann-Whitney U-test); for the novel-voice group $p <$ 0.001 and df = 22 (two-tailed t-test for independent samples). These results are, of course, highly significant.

4. DISCUSSION

The present study demonstrates quite clearly that the acoustic similarity of the voices involved plays a crucial role in shadowing experiments. This does not come as a surprise in view of the fact that Hartmann and Johnson [5] have shown dissimilar non-speech signals to be segregated more easily than similar ones. Our results prove the same to be the case for speech material, even though the overall frequency ranges for both speakers show a large overlap, of course. Thus, voice similarity should be controlled for in future stream segregation experiments.

A question which we will address in future work is whether this effect is only produced by physical dissimilarity or whether it can also be evoked by linguistic dissimilarity, e.g. by speakers who exhibit similar voice characteristics but differ in regional dialect.

Another thing to be learned from this experiment is that the notion of implicit knowledge ought to perhaps be reconsidered. It looks as if various strategies and their success rates have to be distinguished: Did the participant think that he or she knew the voice and therefore try to identify the speaker? If so, at which point in time was that process completed and was the attempt successful? It seems that in the study by Newman and Evers [6] as well as in the present one, a large amount of variability is to be encountered within the implicit-knowledge groups. This is demonstrated by the extremely high standard deviations. Even though some participants correctly guessed the speaker's identity this did not improve their performance to the level of the explicit-knowledge group in [7] and neither did it create a significant difference between the two groups in the present study. Since only one participant remained unaware of the speaker's identity throughout the experiment, all the others having identified him correctly after the first minute, attempts to recognize the speaker (and thus expending processing abilities) can hardly serve as an explanation for this finding. (Incidentally, the person who never identified the target speaker did make more errors than the rest of the group.) Thus it would seem that in future research, implicit knowledge groups will have to be subdivided according to individual processing strategies.

5. REFERENCES

[1] Birkett, P.B., Hunter, M.D., Parks, R.W., Farrow, T.F., Lowe, H., Wilkinson, L.D., Woodruff, P.W. 2007. Voice familiarity engages auditory cortex. *Neuroreport* 18, 1375-1378.

[2] Cherry, E.C. 1953. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America* 25, 975-979.

[3] Craik, F.I.M., Kirsner, K. 1974. The effect of speaker's voice on voice recognition. *Quartely Journal of Experimental Psychology* 26, 274-284.

[4] Darwin, C.J. 2008. Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society Britain* 363, 1011-1021.

[5] Hartmann, W.M., Johnson, D. 1991. Stream segregation and peripheral channelling. *Music Perception* 9, 155-184.

[6] von Hirschhausen, E. 2009. *Die Leber wächst mit ihren Aufgaben* (9th ed). Reinbek: Rowohlt.

[7] Newman, R.S., Evers, S. 2007. The effect of talker familiarity on stream segregation. *Journal of Phonetics* 35, 85-103.

[8] Yarmey, A.D., Yarmey, A.L., Yarmey, M.J., Parliament, L. 2001. Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology* 15, 283-299.