INTERSPEECH 2009 BRIGHTON

Self-voice recognition in 4 to 5-year-old children

Sofia Strömbergsson

Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

sostr@csc.kth.se

Abstract

Children's ability to recognize their own recorded voice as their own was explored in a group of 4 to 5-year-old children. The task for the children was to identify which one of four voice samples represented their own voice. The results reveal that children perform well above chance level, and that a time span of 1-2 weeks between the recording and the identification does not affect the children's performance. F0 similarity between the participant's recordings and the reference recordings correlated with a higher error-rate. Implications for the use of recordings in speech and language therapy are discussed.

Index Terms: human speech perception, speaker identification, children.

10.21437/Interspeech.2009-62

1. Introduction

To many people, the recorded voice often sounds unfamiliar. We are used to hearing our voice through air and bone conduction simultaneously as we speak, and as the recorded speech lacks the bone conduction filtering, its acoustic properties are different from what we are used to [1]. But even though people recognize that the recorded voice sounds different from the voice as we normally hear it, people most often still recognize the recording as the own voice. In a recent study on brain hemisphere lateralization of self-voice recognition [2], adult subjects were given a task of classifying recorded one-word utterances as Self-Voice, Familiar-Voice or Unfamiliar-Voice. A mean accuracy of 95% showed that adults rarely mistake their own recorded voice for someone else's voice.

Although there have been a few studies on adults' perception of their own recorded voices, the literature on children's self-perception of their recorded voices is scarce. However, children's ability to recognize the voices of other speakers has been studied to some extent. In an investigation of pre-school aged children's ability to identify the voices of more and less familiar cartoon characters [3], even the youngest children's performance was significantly better than chance. Children's ability to recognize familiar voices was also explored in [4], where 4 to 5-year-old children were asked to identify the voices of their classmates. Here, large variation in the children's response accuracy was found, with some children performing at adult levels. Both of these studies show that children can identify recorded voices of other speakers, although their performance might not be quite at adult levels.

Different characteristics of the stimuli used in speaker identification tasks have been found to influence the subjects' performance. One important feature is the length of the stimuli; a larger number of phonemes has been reported to have a bigger impact than longer overall duration [5]. Another feature is the similarity between the voices that are to be identified. If only spectral characteristics (including fundamental frequency) are considered, a difference of 11%-16% has been reported as the minimal difference between two utterances for children to perceive the voices as belonging to different speakers [6]. A recent study revealed that the influence of speaking rate on speaker similarity judgment seems to increase after the age of 8 or 9 [7]. There are obviously other acoustic features that contribute to the perception of two recorded voices belonging to the same or two different speakers, e.g. regional accent. However, such features are often more difficult to quantify than speaking rate and spectral features, and their influence on children's judgment of speaker similarity and speaker discrimination remains relatively unexplored.

In [8], children and adolescents (age 7-14) with deviant speech production of /r/ were recorded when pronouncing words containing /r/. The recordings were then edited so that the /r/ sounded correct. A recording in the listening script prepared for a particular child could thus be either an original recording or a "corrected" recording, spoken either by the child himself/herself or another speaker. The task for the children was to judge both the correctness of the /r/ and the identity of the speaker. One of the findings in this study was that the children had difficulty identifying the speaker as himself/herself when hearing a "corrected" version of one of their own recordings. The author speculates that the editing process could have introduced or removed something, thereby making the recording less familiar to the speaker. Another confounding factor could be the 1-2 week time span between the recording and the listening task; this could also have made the task more difficult than if the children had heard the "corrected" version directly after the recording. Unfortunately, no studies of how the time span between recording and listening might affect children's performance on speaker identification tasks have been found, and any effects caused by this factor remain unclear.

Of the few studies that have been done to explore children's perception of recorded voices – of their own recorded voice in particular – many were done over twenty years ago. Since then, there has been a considerable increase in the number of recording devices that can potentially be present in children's environments. This strongly motivates renewed and deeper exploration into children's self-perception of their recorded voice. If it is found that children indeed recognize their recorded voice as their own, this may have important implications for the use of recordings in speech and language intervention.

1.1. Purpose

The purpose of this study is to explore pre-school aged children's ability to recognize recordings of their own voice as their own, and whether this ability is affected by the time span between the recording and the listening. The research questions are:

1. Are children able to recognize their own recorded voice as their own, and identify it when presented in comparison with 3 reference child voices?

Copyright © 2009 ISCA

2. Is this ability affected by the time span between recording and listening?
3. Is the children's performance affected by the acoustic similarity between their voice and the reference child voices?

It is hypothesized that the children will perform better than chance, and that they will perform better when listening immediately after the recording than when listening 1-2 weeks after the recording. Moreover, it is hypothesized that children's performance is affected by the degree of acoustic similarity between a participant's recording and that of a reference child.

2. Method

2.1. Participants

27 children with Swedish as their mother tongue, and with no known hearing problems and with no previous history of speech and language problems or therapy were invited to participate. The children were between 4 and 6 years old, ranging from 4;3 (years;months) to 5;11 (M = 5;3, SD = 6.7 months). Only children whose parents did not know of or suspect any hearing or language problems in the child were invited. All children were recruited from pre-schools in Stockholm. Consent forms were used which complied with Swedish ethical guidelines for subject participation.

2.2. Material

A recording script of 24 words was constructed (see Appendix). The words in the script all began with /tV/ or /kV/, and all had primary stress on the first syllable.

Three 6-year old children (two girls and one boy, included by the same criteria as the children participating in the study) were recorded as references. None of the reference children were known to the children in the test groups.

2.3. Recording/Identification procedure

A computer program was used to present the words in the scripts in random order. For each word, the program first played a reference voice (adult) that read a target word, while displaying a picture that illustrates the word. Then, the child's production of the same word was recorded (with the possibility of listening to the recording and re-recording until both child and experimenter were satisfied). Last, the child's production was presented together with the 3 reference children's productions of the same word, in random order, letting the child select one of these as his/her own (see Fig. 1).

In both test sessions, the children were fitted with a headset and the experimenter with headphones to supervise the recordings. The children were instructed to select the character they believed represented their own voice by pointing at the screen; the actual selection was managed by the experimenter by mouse clicking. The children were given two introductory training items, to assure understanding of the task.

In the first test session, the children performed both the recording and the voice identification task. For the recordings, all children were instructed to speak with their normal voice, and utterances were re-recorded until both child and experimenter were satisfied. In the second test session, after a period of 1-2 weeks, the children performed only the identification task. Apart from general encouragement, the experimenter provided no feedback regarding the children's performance during the voice identification task. All actions – recording, listening and selecting – were logged by the computer program.

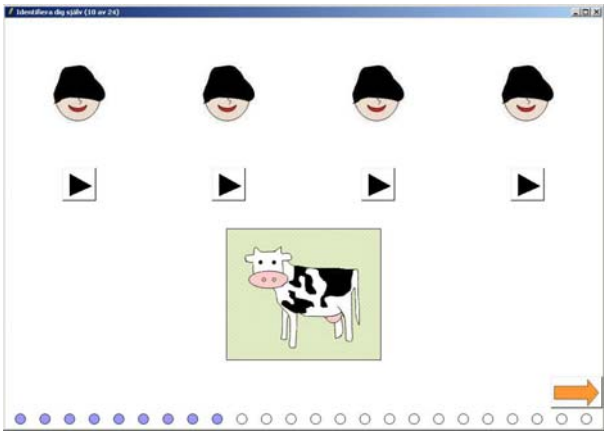

Figure 1: *The listening/identification setup*.

2.4. Voice similarity measurements

Automatic alignment of the speech signal to phonemes was attempted on all 24 words in the script, using NALIGN [9]. For unknown reasons, automatic alignment failed systematically on the word "tub" (*tube*), and therefore this word was discarded from further calculations. Based on the alignments, average fundamental and formant frequencies of each recording were extracted automatically (on voiced parts of the utterances) using the Snack tools F0 and FORMANT [10].

For each of the children's recordings, the absolute difference between the child's F0 and the F0 of the corresponding reference recordings was calculated. The minimal F0 difference (i.e. the absolute difference between the child's F0 and the most similar reference F0) was used as a simple measure of the F0 similarity between the child's recording and the reference recordings. Similarly, the minimal Euclidean distance between the average formant frequencies (F1-F4) of a child's recording and the formant frequencies of the corresponding reference recordings was used as a measure of the spectral similarity between the child's recording and the reference recordings. For each child, the average F0 was calculated and compared to the three reference speakers; the difference between the child's average F0 and the most similar reference F0 was used as a measure of the child's average F0 similarity to the reference speakers. A child's average formant similarity was calculated analogously.

The speaking rate (phonemes per second) was also calculated for all recordings. Initial and final silent parts were excluded before utterance duration was calculated. The minimal difference in speaking rate (i.e. the absolute difference between the child's speaking rate and the most similar reference speaking rate for the same utterance) was used as an additional measure of how similar a child's recording was to the reference recordings. For each participant, average speaking rate similarity to the reference speaking rates was calculated in analogy with the calculation of average F0 similarity described above.

3. Results

Table 1 displays the mean correct own-voice identification for all 27 children on both test occasions. The standard deviation

reveals a quite large variation within the group; the performance varies between 4 and 24 in both the first and the second test. However, the mean correct responses on both tests reflect high average performance results. A closer look at the individual results reveals that two children performed at chance level (or worse), while 7 children (26% of the children) performed with more than 90% accuracy.

A paired samples t-test reveals no difference between performance on the first test and on the second test ($t(26) = 1.517$, $p = 0.141$).

Table 1. *Mean correct responses on the first and second test (max score/test = 24)* .

	Mean correct responses	Std. deviation
First test	18.8 (78.3%)	5.5
Second test	17.9 (74.6%)	6.2
Total	36.7 (76.5%)	11.3

A one-way ANOVA, with the results from the first test as the dependent variable and the number of phonemes as a fixed factor was conducted. This showed that the number of phonemes in a word did not influence the error-rate for that particular word ($F(3,617) = 0.130$, $p = 0.942$). The same was done with the results from the second test as the dependent variable; neither here did the number of phonemes in the stimuli affect the children's performance ($F(3,617) = 0.181$, $p = 0.909$).

Potential effects of acoustic similarity between a child's recording and the most similar reference child recording were examined by a binary logistic regression. Similarity in F0 between the child's recording and a reference recording was found to be a highly significant predictor of the children's response accuracy on the first test ($B = 0.010$, $Exp(B) = 1.010$, $p = 0.004$) and on the second test ($B = 0.016$, $Exp(B) = 1.016$, $p < 0.001$). Thus, with every Hertz increase of the difference between the F0 of a child's recording and the F0 of the most similar reference recording, the odds that the child's answer on the second test will be correct increase by a factor of 1.016. Spectral similarity, which strongly correlates with the F0 similarity, was found to be a highly significant predictor of the children's response accuracy only on the second test ($B = 0.001$, $Exp(B) = 1.001$, $p = 0.006$), whereas its predictive power was not significant for the result in the first test ($B = 0.001$, $Exp(B) = 1.001$, $p = 0.056$). Similarity in speaking rate did not have any significant effect on the children's response accuracy neither on the first test ($B = 0.039$, $Exp(B) = 1.040$, $p = 0.709$) nor on the second test ($B = 0.056$, $Exp(B) = 1.058$, $p = 0.578$).

As alternative measures of F0 and spectral similarity, difference in semitones, and Mahalanobis distance between subject and reference MFCCs (as extracted by the Snack tool SPEATURES [10]) were calculated. Difference in semitones was found to be the only significant predictor, increasing the chances of a correct answer on the second test ($B = 0.078$, $Exp(B) = 1.082$, $p = 0.001$).

Correlations between a child's average acoustic similarity to the reference voices and the child's results on the first and second test were also explored. A weak but significant correlation was found between the children's average F0 similarity to the reference voices and their results on the second test ($r(25) = 0.424$, $p = 0.027$). The correlation between the children's average F0 similarity to the reference voices and their results on the first test was even weaker and only approached significance ($r(25) = 0.338$, $p = 0.085$). No other significant correlations were found.

4. Discussion

The high average performance rates confirm that 4 to 5-year-old children are indeed able to recognize their recorded voice as their own. However, large variation was found among the children, with a few children performing at chance level (or worse) and more children performing with 90% accuracy or more.

No significant difference was found between the children's performance on the first and the second test. The hypothesis that children would perform better when listening immediately after the recording than after a period of 1-2 weeks could thus not be confirmed. The suggested interpretation of the results in [8], that children's difficulties to identify themselves as the speaker in "corrected" versions of their own recordings could be explained by the time span between the recording and the identification task, could thus not be supported. However, the children in [8] were all older than the children in this study, and it cannot be excluded that children of different ages react differently to the time between the recording and the identification task. As many studies have indicated an age-effect in children's ability to recognize and identify familiar voices (e.g. [3], [4], [7]), there is an obvious need to study potential developmental effects on the ability to recognize the recorded voice as one's own.

From the measures of acoustic similarity between the children's recordings and the reference recordings that were used in this study, F0 similarity proved to have most influence on the children's results, both on the level of the individual recording and on the level of speaker averages. So, the bigger the difference in F0 between a child's recording and the reference recordings, the more likely it is that the child would identify this recording accurately. Moreover, the bigger the difference between a child's average F0 and the average F0 in the reference voices, the more likely it is that the child would have a higher score on both the first and the second test. Of the four children with the most similar average F0 to the reference voices (F0 difference ranging from 9 to 16 Hz), two were also found in the group of three children with the total results of 15 or below. The other two, however, had results at 85% total accuracy. So, although F0 similarity has some effect on the children's response accuracy, it can not alone explain why some children have more difficulties identifying their recorded voice than others.

A closer look at the cases where the children selected a reference recording as their own voice revealed that the acoustically most similar reference recordings were actually not preferred over acoustically less similar reference recordings. Thus, although F0 similarity between the child's own recorded voice and the recorded reference voices often leads to an incorrect selection, the selection is not necessarily the reference voice that is most similar to the recording of the child's own voice. More detailed acoustic analysis of the similarity between recordings might provide insights for understanding the children's selection behavior.

The fact that similarity in speaking rate did not have any effect on the children's response accuracy should not be surprising considering the finding in [7], that speaking rate similarity becomes more influential after the age of 8 or 9 in judging speaker similarity. Again, this motivates the exploration of developmental aspects of children's ability to identify the recorded voice as their own.

The large variation found among the children could be due to differences in attention, concentration or understanding of the task, but may also be explained by a difference in aptitude for the task at hand. A closer inspection of the recordings and results of the three children with the worst results (with a total score of 15 or below) revealed that two of these children actually produced slightly deviant speech (despite their parents' assurance that their children had normal speech and language development). This was also noted by the experimenter at the time of the recordings, judging both from the recordings and the children's spontaneous speech. One of the children (a girl aged 5;8) produced [j] for /r/. Another child (a boy aged 5;4) exhibited the same /r/-deviation, together with a few cluster simplification patterns, such as [tɑːva] for "tavla" (*picture*). For the third of these children (a boy aged 4;4), and for all of the other children included in the study, no speech production deviations were noted or could be detected in the recordings. This might suggest a correlation between deviant speech production and difficulties of recognizing the recorded voice as one's own. However, a contradictory example was also found that had to be excluded from the study. Dentalisation (i.e. systematic substitution of [t], [d] and [n] for /k/, /g/ and /ŋ/, respectively) was noted for one girl who could not participate for a second test, and who was therefore excluded from this study. Interestingly, this girl scored 23 of 24 on the first test. These single cases do certainly not present a uniform picture of the relation between deviant speech production and the ability to recognize the recorded voice as one's own, but rather illustrate the need for further investigation of this relation.

In this study, the children's speech production was only controlled to the extent that the experimenter instructed the children to speak with their normal voice, both when introducing the children to the task and whenever the experimenter judged that the child was somehow "playing" with his/her voice. However, some children tended to be more playful than others, and it is unlikely that all recordings reflect the children's normal speech behavior (whatever that is). Although this might certainly have an impact on the results – the children might recognize their speaking behavior rather than their own voice – this would have been difficult to avoid. Moreover, considering that speech play is often encouraged in clinical settings, one could argue that this is also ecologically valid. The results in this study give support to the use of recordings in a clinical setting, e.g. when promoting awareness in the child of deviations in his/her speech production. An example of an effort in this direction is presented in [8], where children were presented with original and "corrected" versions of their own speech production. However, the great variation between children in their ability to recognize their recorded voice as their own requires further exploration, as does the potential developmental changes in this ability.

5. Conclusions

The findings in this study indicate that children as young as 4-5 years old can indeed recognize their own recorded voice as their own. However, there is a large variability among the children; a few children perform at chance level or worse, and many children perform with more than 90% accuracy. Furthermore, children's performance is not significantly affected by a time span of 1-2 weeks between recording and identification. And finally, the more similar the F0 in a recording of a child is to the F0 of another child, the more likely the child is to select the wrong recording as his/her own.

6. Acknowledgements

This work was funded by The Swedish Graduate School of Language Technology (GSLT). Our deepest thanks go to all children participating in this study, and to their teachers and parents for their support in the recruitment process.

7. References

[1] Maurer, D. and Landis, T., "Role of bone conduction in the self-perception of speech", Folia Phoniatrica, 42(5): 226-229, 1990.

[2] Rosa, C., Lassonde, M., Pinard, C., Keenan, J. P. and Belin, P., "Investigations of hemispheric specialization of self-voice recognition", Brain and Cognition, 68(2), 204-214, 2008.

[3] Spence, M. J., Rollins, P. R. and Jerger, S., "Children's Recognition of Cartoon Voices", Journal of Speech, Language, and Hearing Research, 45(1), 214-222, 2002.

[4] Bartholomeus, B., "Voice identification by nursery school children", Canadian Journal of Psychology/Revue canadienne de psychologie, 27(4), 464-472, 1973.

[5] Bricker, P. D. and Pruzansky, S., "Effects of Stimulus Content and Duration on Talker Identification", The Journal of the Acoustical Society of America, 40(6), 1441-1449, 1966.

[6] Cleary, M., Pisoni, D. B. and Iler Kirk, K., "Influence of Voice Similarity on Talker Discrimination in Children with Normal Hearing and Children With Cochlear Implants", Journal of Speech, Language, and Hearing Research, 48(1), 204-223, 2005.

[7] Petrini, K. and Tagliapietra, S., "Cognitive Maturation and the Use of Pitch and Rate Information in Making Similarity Judgments of a Single Talker", Journal of Speech, Language, and Hearing Research, 51(2), 485-501, 2008.

[8] Shuster, L. I., "The perception of correctly and incorrectly produced /r/", Journal of Speech, Language, and Hearing Research, 41(4), 941-950, 1998.

[9] Sjölander, K., "An HMM-based system for automatic segmentation and alignment of speech", Proceedings of Fonetik 2003, 93-96, 2003.

[10] Sjölander, K. The Snack sound toolkit, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden. Online: http://www.speech.kth.se/snack/, 1997-2004, accessed on April 12, 2009.

8. Appendix

	Orthography	Transcription	In English
1)	k	/koː/	*(the letter k)*
2)	kaka	/kɑːka/	*cake*
3)	kam	/kam/	*comb*
4)	karta	/kɑːʈa/	*map*
5)	katt	/kat/	*cat*
6)	kavel	/kɑːvəl/	*rolling pin*
7)	ko	/kuː/	*cow*
8)	kopp	/kɔp/	*cup*
9)	korg	/korj/	*basket*
10)	kula	/kʉːla/	*marble*
11)	kulle	/kɵlə/	*hill*
12)	kung	/kʊŋ/	*king*
13)	tåg	/toːg/	*train*
14)	tak	/tɑːk/	*roof*
15)	tant	/tant/	*lady*
16)	tavla	/tɑːvla/	*picture*
17)	tidning	/tiːnɪŋ/	*newspaper*
18)	tiger	/tiːgər/	*tiger*
19)	tomte	/tɔmtə/	*Santa Claus*
20)	topp	/tɔp/	*top*
21)	tub	/tʉːb/	*tube*
22)	tumme	/tʊmə/	*thumb*
23)	tunga	/tʊŋa/	*tongue*
24)	tupp	/tʊp/	*rooster*