An Interactive Technique for Matching Speaker Identity

Rolf Carlson and Björn Granström*
Department of Speech Communication and Music Acoustics,
KTH, Box 70044, S 100 44 Stockholm, Sweden

ABSTRACT
A speaker recognition experiment is described where subjects are asked to identify one out of twelve unknown speakers from short samples of speech. Results on speed and precision of this task will be presented. The study will be used as a baseline for synthesis of speaker characteristics.

INTRODUCTION
Currently available text-to-speech systems are not characterized by a great amount of flexibility, especially not when it comes to varying the voice. However, there are practical needs for different voices. Text-to-speech systems are now used in many applications which ask for voice variation close to that found in human speakers. This is especially true when the systems are used in communication. Speaker-specific aspects are regarded as playing a very important role in the acceptability of synthetic speech (Carlson, Granström & Karlsson, 1991). It is currently an important scientific challenge to formulate human speech variability in explicit models. Thus, a significant ambition in speech synthesis research is to model speech on a global level, allowing changes of speaker characteristics and speaking style. It would be helpful if the speaker characteristics could be described by a limited number of parameters. Only a small number of sentences might in this case be needed to adjust the synthesis to one specific speaker. The needs in speech synthesis research and speech recognition research are very similar in this respect (Blomberg 1991, Ström 1994.) As a starting point for this work we wanted to study how well listeners can judge speaker identity from short samples of human speech.

TEST PROCEDURE
To establish a baseline for speaker recognition, a test procedure was established. Different voices have been collected as part of the "Waxholm" speech database (Blomberg et al. 1993) . In this database, collected as part of a Wizard of Oz recognition experiment, each subject uttered a set of fixed sentences varying in length and linguistic complexity. From this database 12 speakers were selected. All were male speakers without strong regional accents. All speakers were unknown to the subject group, that consisted of 48 students taking the speech communication class at KTH. The test task was to match the voice of a test sentence to one of 12 voices. The test was run individually as an interactive computer-controlled task for 15 minutes. In Figure 1 the procedure is illustrated. The figure resembles the screen display seen by the subject. By clicking the mouse in the central square the subject could listen to the test sentence. By clicking the 12 peripheral squares the subject could listen to the 12 different voices represented in the test. These 12 reference samples were the same sentence uttered by the 12 speakers. During the test session the location of the reference voices were kept the same. The subjects were divided in three groups, for which the location of the references were shifted one step (see Figure 1)

---

* Names in alphabetic order

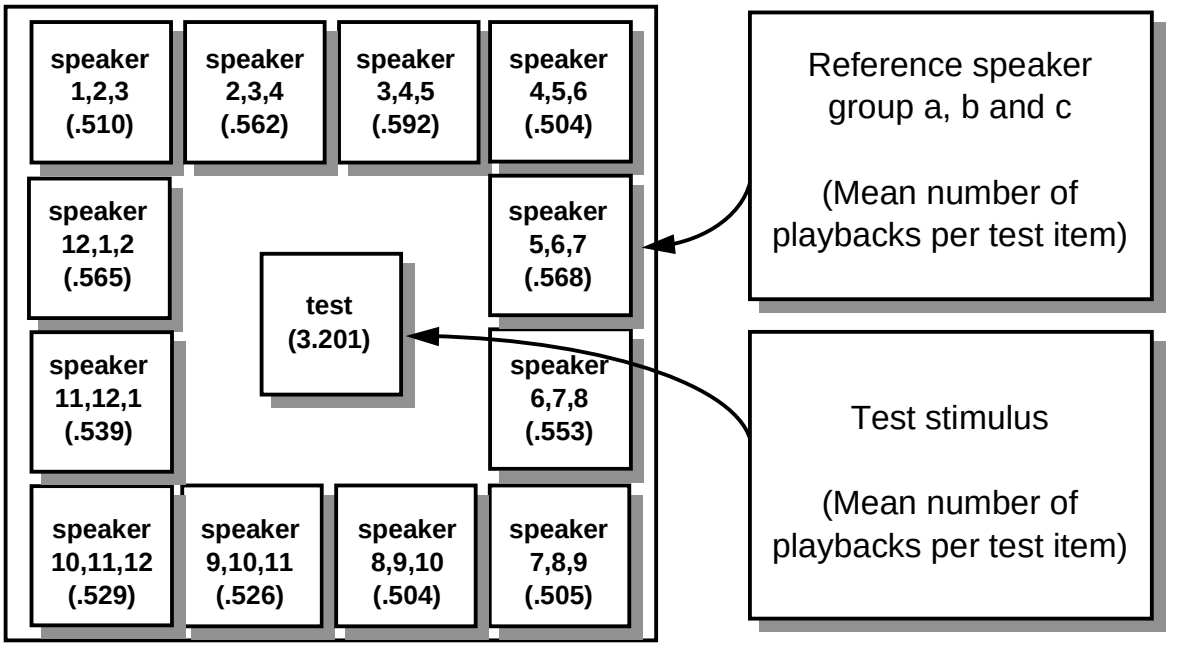

Figure 1. Graphic display presented to the subjects. Reference speakers are positioned along the sides of the main square and the test speaker in the middle.

Reference sentence:	Waxholm ligger i Stockholms skärgård.
Test sentences:	1	Ja.
	2	Lila stolar bärs in i salen.
	3	Det var kyligt i luften och stjärnorna skimrade.
	4	Lediga och utvilade tittade dom på föreställningen i en timme.
	5	Sprakande fyrverkeripjäser exploderade över oss.
	6	Där kommer nya röda hus att skjuta i höjden.

Table 1. Sentence material used in the experiment.

The test sentences (see Table 1) were randomly drawn from six sentences (not including the reference sentence) uttered by the same 12 speakers. The test sequence was the same for all subjects. All selections on the screen were recorded, along with the selection time.

RESULTS

In Figure 2, the proportion of subjects completing a certain number of test items is displayed. The bold line is the mean result (48 subjects) and the thin lines are the results in the individual groups (16 subjects each). It is obvious that the groups are quite comparable, but that there exists a wide variation in the number of test items covered during the 15-minute test period (12 to 60 items with a mean of 27).

In Figure 1, the mean number of times each square is clicked for one test item is indicated. It is a tendency that the corner squares are listened to less than other squares, possibly because the voices in these positions are more easily remembered. This justifies the rotation between groups that was employed in the test.

In Figure 3, the time-to-decision is plotted as a function of the serial position in the test. The mean number of playbacks used to arrive at this decision is also plotted. It is clear that the subjects used the first item to get familiar with the display. After that a more gradual decrease of the time used (and the number of squares clicked) can be seen. This is

in part due to the inclusion of both fast and slow subjects for the earlier test items. Clear variation between test sentences could be seen. The points indicated by circles in Figure 3 all pertain to the very short utterance "Ja." (yes).

In Table 2 the error distribution is given. The overall error was 11.6%. For the different sentences the error varied between 29% for sentence 1 (Ja.) that alone accounted for more than half of the errors, to 4.2% for the easiest sentence (#2). A smaller variation was found between different speakers.

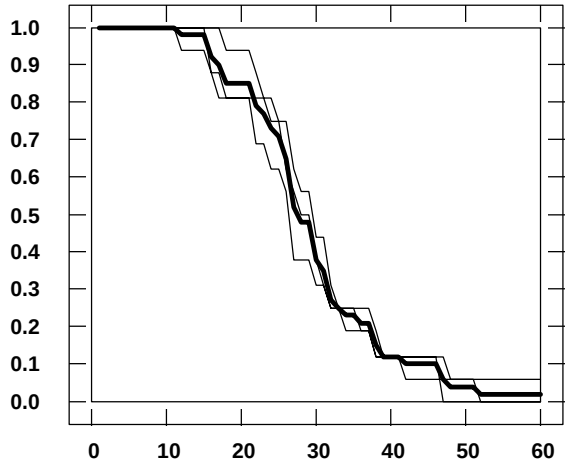

Figure 2. Proportion of subjects completing a certain number of test items. The bold line is the mean result (48 subjects) and the thin lines are the results in the individual groups (16 subjects each).

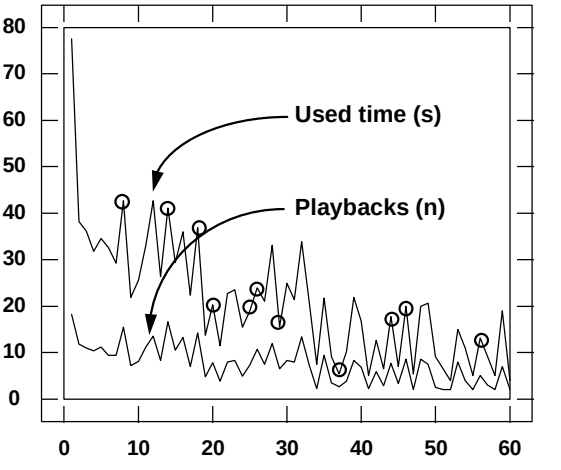

Figure 3. Time to decision as a function of the serial position in the test and number of listened utterances for each test item

sentence	correct	error	% error
1	203	83	29.0
2	296	13	4.2
3	186	13	6.5
4	140	15	9.7
5	245	21	7.9
6	155	15	8.8
total	1225	160	11.6

speaker	correct	error	% error
1	134	21	13.5
2	88	4	4.3
3	119	13	9.8
4	46	9	16.4
5	115	32	21.8
6	171	15	8.1
7	85	15	15.0
8	121	20	14.2
9	78	8	9.3
10	128	9	6.6
11	64	8	11.1
12	76	6	7.3
total	1225	160	11.6

Table 2. Number of correct and incorrect identifications, according to speaker and sentence.

Since the sentences and speakers were randomized in the test, and the subjects were engaged for a fixed duration, we observe great variation in the the number of trials for each sentence and speaker. This is especially true in the speaker statistics.

Figure 4 illustrates another aspect of the data. The time for wrong identification is plotted against time for correct identification. It is displayed for each test item that created an error. The result shows a tendency that it takes longer time to give a wrong response compared to a correct one.

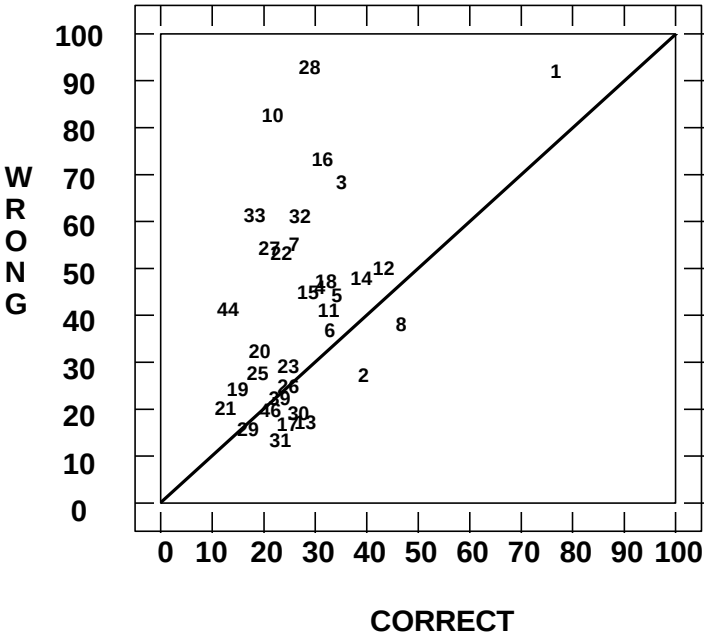

Figure 4. Mean time (in seconds) for arriving at a correct or wrong identification. Numbers refer to the sequential position in the test.

FINAL REMARKS

The result demonstrates that it is possible for subjects to quite accurately match unknown voices, speaking different sentences, if the test sentences are reasonably long. Listeners differ widely in this ability.

We intend to use this material in a synthesis experiment, where we will model the individual speaker characteristics. At the meeting, some of these results will be demonstrated.

ACKNOWLEDGMENT

This work has been supported by The Swedish National Language Technology Program. We want to extend out thanks to Mats Båvegård and Roger Lindell, who implemented the computerized test and to Johan Bertenstam, Joakim Gustavsson, Anders Roxström, Håkan Melin and Nikko Ström that helped in administering the test.

REFERENCES

Blomberg, M. (1991): "Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references," Speech Communication, Vol. 10. pp 453-462.

Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. (1993): "An experimental dialog system: WAXHOLM," Proceedings of Eurospeech '93. pp 1867-1870.

Carlson, R. Granström, B. and Karlsson, I. (1991), "Experiments with voice modeling in speech synthesis," Speech Communication 10, pp 481-489.

Ström, N. (1994): "Experiments with a new method for fast unsupervised speaker adaption in continous speech recognition," To be presented at this meeting.