Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

The Effect of Target/Non-Target Age Difference on Speaker Recognition Performance

George Doddington

george.doddington@comcast.net

Abstract

The very large set of trials in the SRE10 extended evaluation [1] provides opportunity to study the effect of various factors on speaker recognition performance. This paper addresses the issue of age difference between target and non-target speakers and shows that false alarm probability is reduced substantially as the age difference increases. False alarm probability is significantly reduced for age differences of as little as five years, with an order of magnitude reduction in P_{FA} for age differences of forty years or more, depending on the system being measured and the test condition.

1. Introduction

Understood but not often voiced, the performance of a speaker recognition system is affected by the demographics of the speaker population that the application deals with. One such factor is speaker sex, which NIST has dealt with in two ways. First, NIST does not include cross-sex trials in their performance analyses, even though doing so would show better performance. Second, NIST presents results separately for men and women, because the performance for men is different from and typically better than that for women.

Another factor is speaker age, studied in this paper. Speaker age is addressed by analyzing how the age difference between target and non-target speakers affects the performance of a speaker recognition system.

2. The Data

The data used were SRE10 extended evaluation [1] results submissions from four participating sites. The trials were analyzed separately for the two conditions with the greatest number of trials, namely for interview speech and for conversational telephone speech.

Table 1 SRE10 extended data speaker and trial statistics

Condition #	sex	# of speakers	# of target trials	# of non-target trials
2 - interview speech	female	229	8152	1,573,948
	male	196	6932	1,215,586
5 - telephone conversations	female	199	3704	233,077
	male	172	3465	175,973

The distribution of ages in the SRE10 evaluation corpus is shown in figure 1.

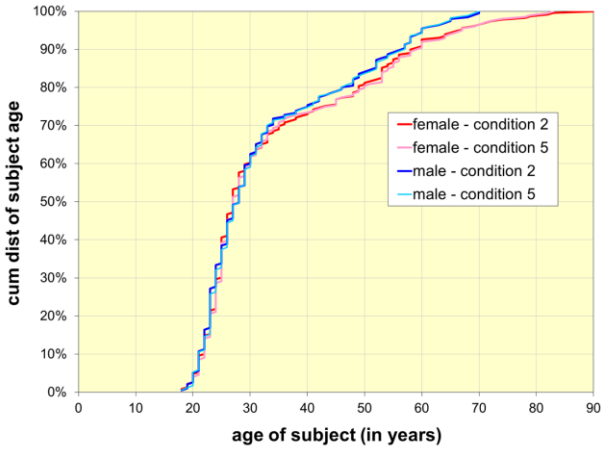

Figure 1 The distribution of speaker ages as a function of sex and test condition.

3. The Method

To determine the effect of age differences between the target speaker and non-target speakers, six different age-difference categories were defined. The age-difference ranges for these categories were chosen so that the number of target speakers in each category was approximately the same for all categories. These age-difference categories and related counts of speaker pairs are shown in table 2.

The effect of the age difference was represented in terms of the false alarm probability at a given miss probability. Although a P_{Miss} value of 10% is commonly used [2], this study used a P_{Miss} value of 1% instead, in order to avoid unreliable estimates of P_{FA} due to a small number of non-target errors, zero in some cases. P_{FA} was computed separately for men and women and for condition 2 and 5.

Care was taken to balance the contribution from different non-target speakers and from different target speakers. To do this, the following procedure was used for each age-difference category:

1. A DET curve was created for each target/non-target speaker pair. This was done so that the unique statistics of each speaker pair would have equal weight in creating an overall DET curve.
2. Then an average DET curve for each target speaker was created by averaging together all of the speaker pair DET curves created in step 1.[1] This was done in order to weight equally the contribution of each non-target speaker.

[1] An "average" DET curve is a DET curve formed by averaging, as a function of score, the P_{Miss} and P_{FA} of the constituent DET curves in the set to be averaged.

3. Finally an overall average DET curve was created by averaging together all of the target speaker DET curves created in step 2. This was done in order to weight equally the contribution of each target speaker.

Note that not all categories contain data from exactly the same set of target speakers. This is because some target speakers have no non-target trials in one or more age-difference categories due to a lack of non-target cohorts in those categories. To eliminate this source of variance, the set of target speakers used in the following analysis was pruned to keep only those target speakers that had non-target cohorts in all age-difference categories. This reduced the number of target speakers used in the age-difference analysis to that shown in table 2.

Table 2 target and non-target speaker statistics used for the age-difference performance analysis

Condition #	**2** - interview speech		**5** - telephone conversations	
sex	female	male	female	male
# of targets	212	164	173	143
average # of non-targets per target				
age diff **0-4** yrs.	80.5	74.0	77.0	69.5
age diff **5-9** yrs.	43.6	47.1	42.9	43.6
age diff **10-19** yrs.	35.0	31.4	27.3	27.8
age diff **20-29** yrs.	30.1	21.9	18.9	19.0
age diff **30-39** yrs.	23.7	22.8	23.2	21.6
age diff **40-99** yrs.	21.0	10.9	21.7	10.4

4. The Results

Performance is represented in terms of the false alarm rate at a given miss rate. This performance measure has been used in IARPA's BEST program [2], with the given target miss rate being 10%. This typically produces very small false alarm rates, and in the corpus studied a miss rate of 10% results in no false alarms for some age-difference categories and some systems. Unfortunately, a P_{Miss} value of 10% is too small to be reliably indicative for the purposes of this study. So the benchmark value for P_{Miss} was chosen to be 1%. Given this miss rate, figures 2 through 5 illustrate how false alarm performance improves as the age difference between target and non-target increases. For condition 2, P_{FA} for the 5-9 year age-difference category is about 7-12 percent less than for the 0-4 year age-difference category, and P_{FA} for age differences of 40 or more years is reduced by a factor of 2-9, depending on the system. For condition 5 the P_{FA} contrast between different age categories is even more striking, with P_{FA} reductions of more than an order of magnitude for some systems.

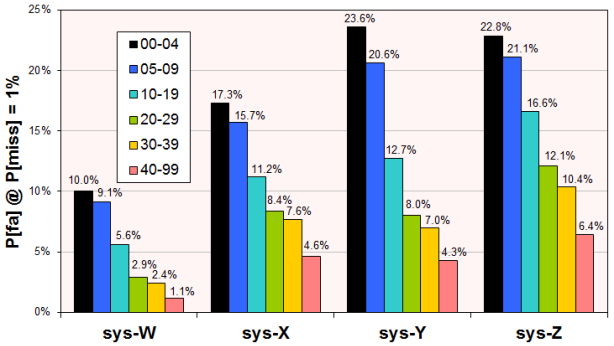

*Figure 2 Non-target false alarm performance for **condition 2** as a function of the target/non-target age difference for **female speakers***

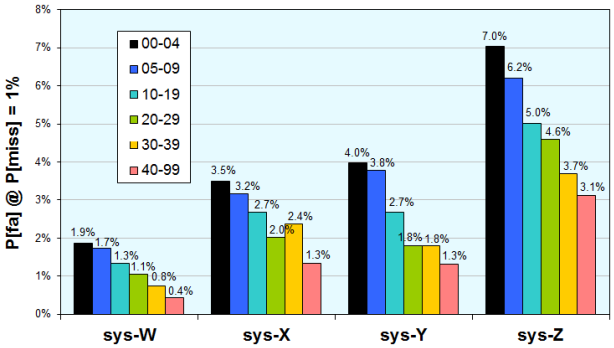

*Figure 3 Non-target false alarm performance for **condition 2** as a function of the target/non-target age difference for **male speakers***

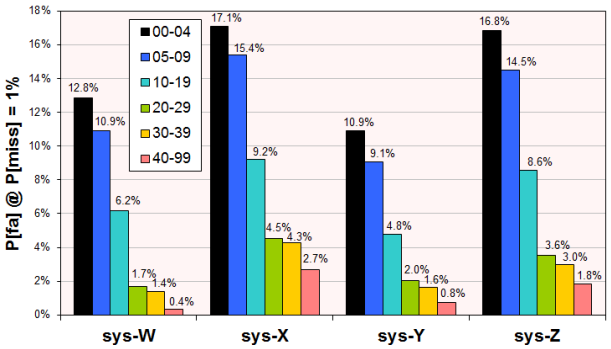

*Figure 4 Non-target false alarm performance for **condition 5** as a function of the target/non-target age difference for **female speakers***

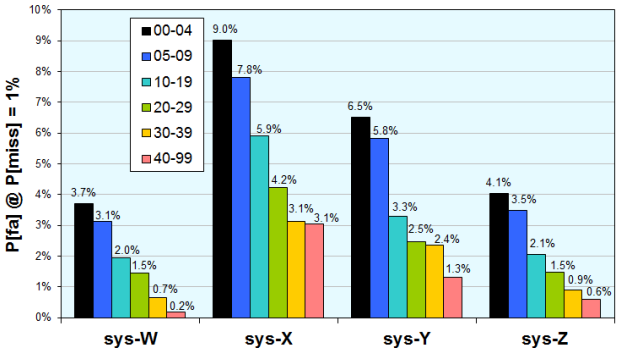

*Figure 5 Non-target false alarm performance for **condition 5** as a function of the target/non-target age difference for **male speakers***

While the differences between P_{FA} for different categories appear to be quite significant, it would be good to verify this by computing confidence intervals on P_{FA}. To this end, 90% confidence curves for P_{FA} are shown in Figure 6, which were computed assuming that the DET curves for different target speakers are statistically independent. Note that the confidence interval is quite small at 1% P_{Miss} and appears adequate to assert that the performance differences observed for the various age-difference categories are generally statistically significant. Note also that variance attributable target scores is not accounted for. This is because the target scores are identical for all age-difference categories and therefore only the variance of P_{FA} need be accounted for. However this also means that the PFA confidence intervals are valid only for this particular set of target speakers and their particular scores. The confidence intervals would be much larger for random selection of target speakers or target speaker trials.

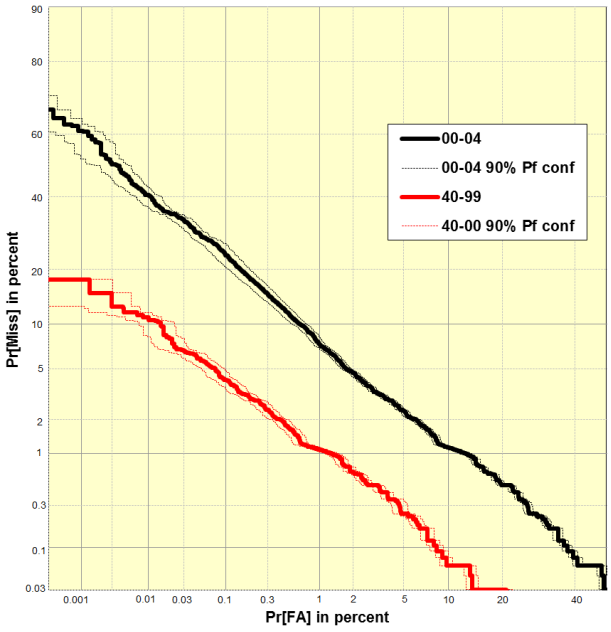

Figure 6 DET curves showing 90% confidence intervals for P_{FA} for the two extreme age-difference categories, for system sys-W *for condition 2. DET curves are for female and male trials combined.*

It should be noted that the number of non-targets for a given target varies and sometimes is very small, small enough to make the resulting average DET curve for that target speaker highly unreliable. In an attempt to improve the quality of the results, a modified analysis was performed that included only those target speakers with at least 8 non-target cohorts. The number of excluded target speakers is shown as a function of age-difference category in figure 7.

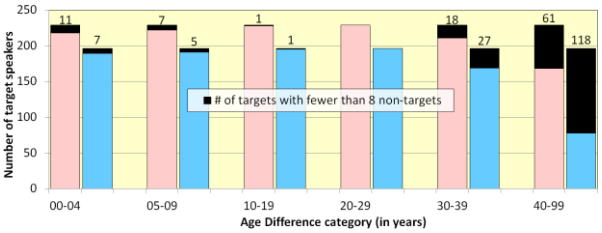

Figure 7 The number of target speakers in condition 2, 229 women (pink) and 196 men (blue), showing the number of targets that have fewer than 8 non-target cohorts, as a function of age-difference category

This requirement that target speakers have at least 8 non-target cohorts results in a very significant reduction of target speakers, because of the requirement that every target speaker be represented in all age-difference categories. The resulting statistics are shown in table 3.

Table 3 target and non-target speaker statistics used for the age-difference performance analysis – target speakers are limited to those with at least 8 non-target cohorts

Condition #	**2** - interview speech		**5** - telephone conversations	
sex	female	male	female	male
# of targets	157	71	133	60
average # of non-targets per target				
age diff **0-4** yrs.	102.1	83.6	93.5	80.8
age diff **5-9** yrs.	46.1	40.7	42.5	36.0
age diff **10-19** yrs.	21.4	23.4	17.6	20.7
age diff **20-29** yrs.	22.2	21.3	17.7	17.8
age diff **30-39** yrs.	24.8	27.4	23.1	25.7
age diff **40-99** yrs.	17.5	11.5	16.5	11.0

The question now is whether the selection of fewer targets with more stable DET curves yields better results. To address this question, figures 8 and 9 plot the standard error of P_{FA} (computed over all target speakers), measured at an overall value of 0.1% P_{FA}, for the two extreme age-difference categories, with each plot showing the standard error for each of the four systems. Note that for most cases the standard error for the limited set of target speakers being averaged is less than that for the full set of target speakers. This suggests that limiting the DET curves being averaged to only those for target speakers with at least 8 non-target cohorts does seem to stabilize the resulting averaged DET curve, at least for the two extreme age-difference cases. Additional support for improved averaging is provided by comparing the DET curves for system sys-W shown in figure 10, produced by averaging the DET curves for all target speakers, with the DET curves in figure 11, produced by averaging the DET curve for only target speakers with at least 8 non-target cohorts. With the limited set of target speakers the DET

curves appear to be better behaved in the low false alarm region.

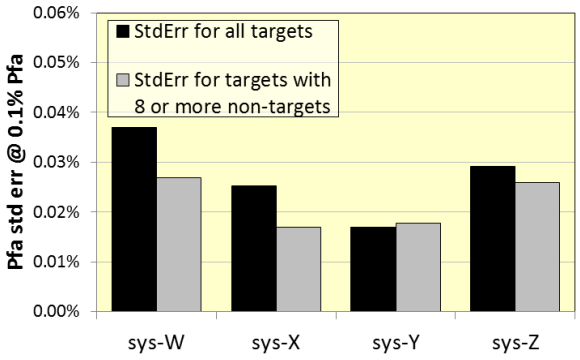

Figure 8 The standard error for P_{FA} at a decision threshold that yields 0.1% P_{FA} for age-difference category 00-04 for condition 2.

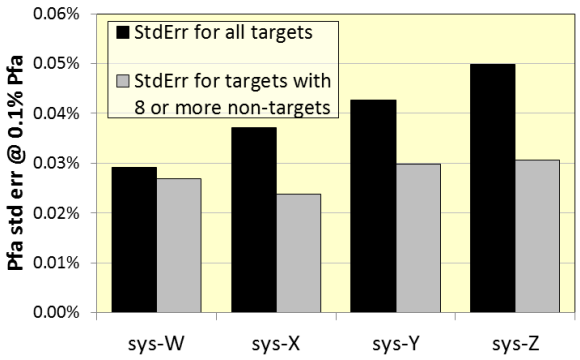

Figure 9 The standard error for P_{FA} at a decision threshold that yields 0.1% P_{FA} for age-difference category 40-99 for condition 2.

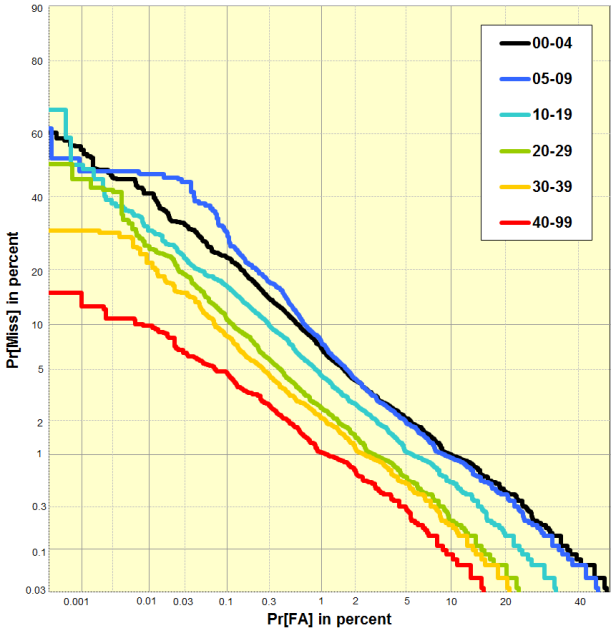

Figure 10 age-difference DET curves for system sys-W *for condition 2, produced by averaging DET curves for all target speakers. DET curves are for male and female speakers combined.*

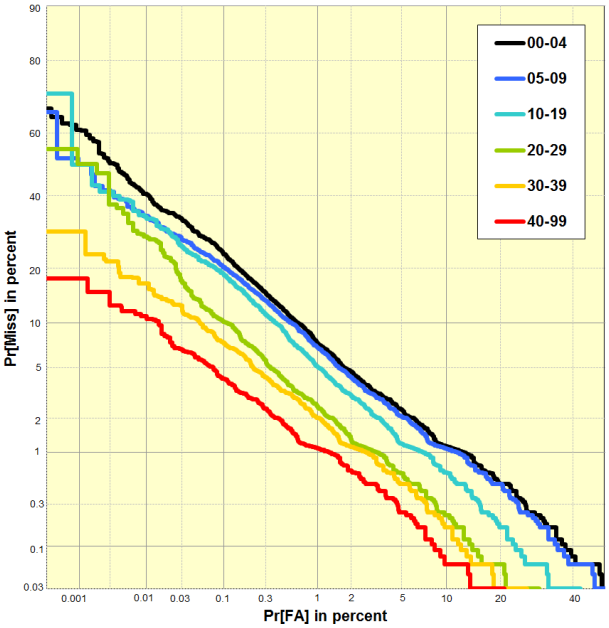

Figure 11 age-difference DET curves for system sys-W *for condition 2, produced by averaging DET curves for only those target speakers with 8 or more cohorts. DET curves are for male and female speakers combined.*

Using the limited set of target speakers with at least 8 non-target cohorts, non-target P_{FA} performance was recomputed, averaging DET curves over just this limited set of targets. These performance statistics are shown in figures 12-15, for comparison with those shown in figures 2-5. There are several surprises in the comparison. First, the limited averaging appears not to have served the objective of providing more reliable performance estimates, at least from the observation that there are more violations of monotonic performance trends with age difference in the limited averages than in the unlimited averages. Second, the two different sets of target speakers produce P_{FA} results that are quite different from each other. This dramatizes the effect of target speaker selection and justifies the requirement that exactly the same set of target speakers be used in all age-difference categories.

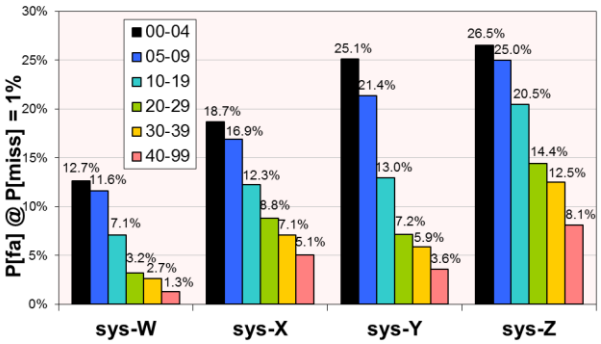

Figure 12 Non-target false alarm performance for ***condition 2*** *as a function of the target/non-target age difference for* ***female speakers****. Target speakers are limited to only those with at least 8 non-target cohorts*

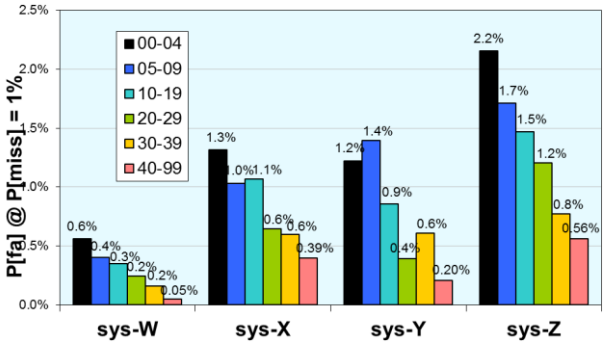

*Figure 13 Non-target false alarm performance for **condition 2** as a function of the target/non-target age difference for **male speakers**. Target speakers are limited to only those with at least 8 non-target cohorts*

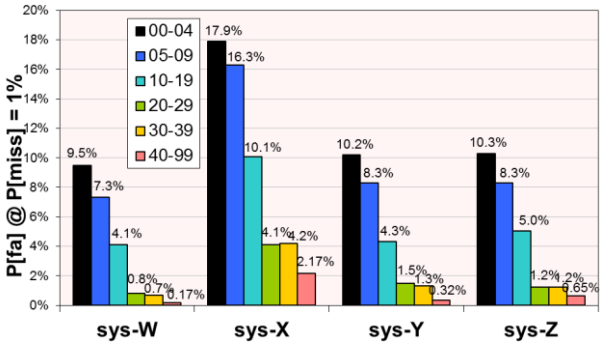

*Figure 14 Non-target false alarm performance for **condition 5** as a function of the target/non-target age difference for **female speakers**. Target speakers are limited to only those with at least 8 non-target cohorts*

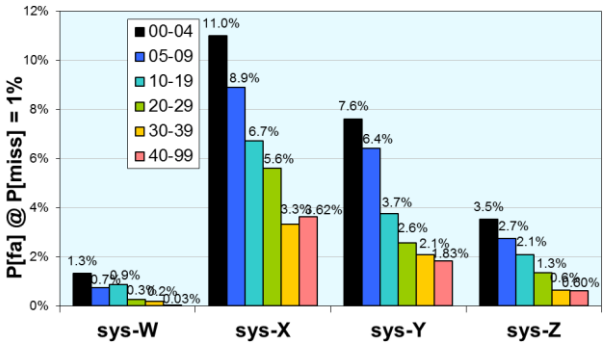

*Figure 15 Non-target false alarm performance for **condition 5** as a function of the target/non-target age difference for **male speakers**. Target speakers are limited to only those with at least 8 non-target cohorts*

5. Causes of Performance Differences

While a number of studies have been conducted related to the effect of ageing on speaker recognition performance, e.g. [3][4], this study is a bit different in that the issue in this study is (non-target) population differences rather than target speaker differences attributable to longitudinal ageing effects. Note that the performance contrasts in this study used exactly the same set of target trials in each age-difference category, and that the threshold setting (to achieve 1% PMiss) was therefore the same for all age-difference categories.

This study did not attempt to understand or probe the factors that underlie age-related performance differences, only to expose and calibrate the differences. That said, there are surely many factors that contribute, including not just physical and physiological factors but also language factors, such as evolving word pattern usage that influences a speaker's idiolect. [5]

6. Summary and Exhortation

This follow-up study to the SRE10 extended data evaluation demonstrates that performance varies dramatically with differences in age between target and non-targets. A very important lesson to be learned is that population demographics are critically important in determining the performance of speaker recognition systems. Age is of course just one of many demographic factors of importance. Some of these are fairly obvious. Some are not. Still others may remain to be discovered. Thus population demographics should be considered with great care in the course of planning a speaker recognition research effort or evaluation project.

7. References

[1] The NIST 2010 speaker recognition evaluation, extended data. See www.nist.gov/itl/iad/mig/sre10.cfm

[2] http://www.iarpa.gov/solicitations_best.html

[3] Lei, Yun / Hansen, John H. L. (2009): "The role of age in factor analysis for speaker identification", In INTERSPEECH-2009, 2371-2374.

[4] Finnian Kelly, Naomi Harte, "Effects of long-term ageing on speaker verification" In *Biometrics and ID Management*, volume 6583 of *Lecture Notes in Computer Science*, pages 113–124. Springer Berlin / Heidelberg, 2011

[5] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," Eurospeech, Vol. 4, pp. 2517-2520, 2001