# Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques

*Nobuaki MINEMATSU*[*†], *Keita YAMAUCHI*[**], *and Keikichi HIROSE*[‡]

[*]Graduate School of Information Science and Technology, University of Tokyo
[†]Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm
[**]Faculty of Engineering, University of Tokyo
[‡]Graduate School of Frontier Sciences, University of Tokyo
{mine,kta-yama,hirose}@gavo.t.u-tokyo.ac.jp

## Abstract

This paper proposes a technique to estimate speakers' perceptual age automatically only with acoustic information of their utterances. Firstly, we experimentally collected data of how old individual speakers in databases *sound* to listeners. Speech samples of approximately 500 male speakers with a very wide range of the real age were presented to listeners, who were asked to estimate the age only by hearing. Using the results, the perceptual age of the individual speakers was defined in two ways as label (averaged age over the listeners) and distribution. Then, each of the speakers was acoustically modeled by GMMs. Finally, the perceptual age of an input speaker was estimated as weighted sum of the perceptual age of all the other speakers in the databases, where the weight for speaker $i$ was calculated as a function of likelihood score of the input speaker as speaker $i$. Experiments showed that correlation was about 0.9 between the perceptual age estimated by the listening test and that estimated by the proposed method. This paper also introduces some techniques to realize robust estimation of the perceptual age.

## 1. Introduction

Although recent advances of speech technologies have brought spoken dialogue systems close to reality, most of the systems equally deal with users even though they naturally have different characters. Further, all the input information from a user is once converted into a sequence of words by a speech recognizer and the user's intention is interpreted by analyzing the sequence. In this case, non-verbal information is usually lost, which can sometimes be a key to detect the hidden intention. Even on the verbal communication only, it may be possible to treat a user differently from another. But in human-to-human communication, it is easily expected that a listener perceives various aspects of a speaker through his/her looks, sounds, smells, feels, and perhaps tastes. Then, the listener often changes his/her strategies to react to the speaker based upon this kind of information, which is frequently transmitted by way of non-verbal media. Man-machine interface has been gracefully changed from CUI to GUI. Researchers on the interface discuss PUI (Perceptual User Interface) as the interface of the next generation[1, 2].

We can easily find many younger children playing or studying with computers these days. The population of elderly people is getting larger year by year, and therefore, more and more elderly people are expected to use computers in their daily lives. These facts mean that spoken dialogue systems should be developed so that they are friendly to all the generations. Although it may be possible to do that only in a unique and universal manner over the generations, dynamic, flexible, and meticulous control of user-interface and dialogue strategies shall be realized if users' age can be correctly estimated[3]. Based on these considerations, in this paper, we focus on automatic estimation of the speakers' age, which can be defined in two ways, biological age and perceptual age. In this paper, the latter is estimated because a listener always uses information of the perceptual age when he/she changes the reacting strategy based on the speaker's age.

## 2. Subjective estimation of speakers' age

### 2.1. Speakers and subjects

In this listening experiment, three large databases were used, which are JNAS (Japanese News Article Sentence) and Senior-JNAS databases and a database of CHILDREN's speech. The first one gave us speech samples of 153 adult male speakers of age ranging from 20 to 60 and, from the second one, 202 male speakers's speech samples were given. Their age varied from 60 to 90. Speech samples of 6 to 12-year-old boys were used from the third one. The number of speakers was 141. As for subjects, 30 university students joined the experiment. Perceptual age was supposed to depend upon the real age of the subjects and all of them were in their early 20's. All the listening and answering were done on web pages for efficient data collection.

### 2.2. Procedures of the listening experiment

The listening experiments were done in a quiet room by using headphones with a fixed volume level. One sentence utterance per speaker was presented to a subject and he/she was required to estimate the age in a range of 0 to 100 by a unit of 1. Order of the presentation was random and correspondence between a sentence and a speaker was changed among the subjects because the linguistic content was supposed to have some effects on the human age estimation. After that, the subject was also asked to select a noise level of the given utterance out of three candidates, low, middle, and high. This is because the children's speech database contained many speech samples with noise. These data will cause some undesired effects on the analysis and speakers with high-level noise will be excluded.

### 2.3. Results of the subjective age estimation

The perceptual age averaged over the subjects is shown for each speaker in Figure 1. Clearly seen, speakers with a wide range of the perceptual age were prepared for this study. Standard deviation was also calculated for each speaker. Although the graph is not shown due to limit of space, as is expected, it is smaller in children's speech and larger in adults' speech. In other words, the perceptual age has a sharp distribution for a younger speaker

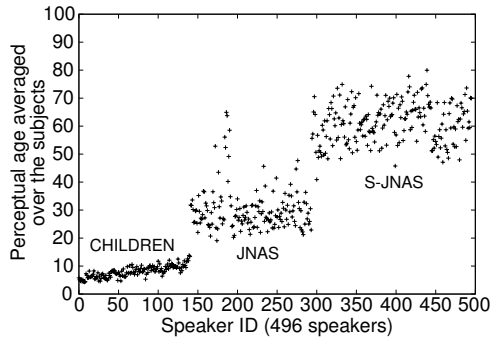


Figure 1: *Perceptual age averaged over the subjects*

Table 1: *Acoustic conditions*

| | |
|---|---|
| training data | JNAS (130 speakers), SJNAS (190 speakers) |
| (60 sec) | CHILDREN (82 speakers) |
| testing data | The same above |
| (5 sec) | The testing was done by cross-validation. |
| sampling | 16 kHz / 16 bit |
| window | 25 msec Hamming window with 10 msec shift |
| preemphasis | $1.0 - 0.97z^{-1}$ |
| parameters | 12MFCC + 12$\Delta$MFCC + $\Delta$Power |
| GMM | 16 mixtures with diagonal covariance matrices |

and a broad one for an older speaker. In Section 4, the perceptual age of a speaker is defined as a label of the averaged age but as a normal distribution with mean and variance in Section 5.

## 3. Acoustic modeling of the speakers

After the listening test, 89 speakers (18 in JNAS, 12 in SJNAS, and 59 in CHILDREN) were excluded due to the high noise level and the remaining 407 speakers were used in the subsequent experiments. Each of the speakers was acoustically modeled as GMM under the acoustic conditions listed in Table 1. For the modeling, all the silent frames were removed from the original frame sequence and a 60 sec sequence out of the remaining was used to model a speaker. Testing was done only with another 5 sec frame sequence for each test speaker.

Previous studies showed that power spectrum in higher frequency bands of elderly speech are reduced compared to those of non-elderly speech[4], that elderly speakers and non-elderly ones can be identified with speaker modeling techniques[5], and that elderly speech recognition performance is improved by adapting acoustic models to elderly speech[6]. These results indicate that spectrum envelopes carry some information on the speakers' agedness. In the following sections, the perceptual age is estimated by using the GMMs of all the speakers.

## 4. Estimation of the perceptual age with its labels

### 4.1. Estimation of the perceptual age through expectation

In this paper, given acoustic observations $o$, the perceptual age (PA) of an input speaker is estimated as an expected value of

$$PA = \frac{\sum_x P(x|o)x}{\sum_x P(x|o)}, \tag{1}$$

where $x$ is a perceptual age value estimated through the listening experiment. Although $\sum_x P(x|o)$ is theoretically constant to be one, it is explicitly written for the later discussion. According to Bayes's Rules, $P(x|o)$ can be re-written as

$$P(x|o) = \frac{P(o|x)P(x)}{P(o)}. \tag{2}$$

If $P(o)$ is considered as constant term, the expectation operation can be done by using $P(o|x)P(x)$ for $P(x|o)$. $P(x)$ is a prior probability distribution and, for example, it corresponds to a distribution of the perceptual age of users of a system. But this distribution strongly depends upon the system and, in this work, an even distribution was assumed as $P(x)$. Finally, the expectation operation can be done by using $P(o|x)$ as weights. It should be noted that $P(o|M_i)$ ($M_i$ is a GMM of speaker $i$) cannot be substituted directly for $P(o|x)$. This is because the perceptual age in the databases has a biased distribution. Namely, the number of elderly people is larger. If $P(o|M_i)$ is used, the estimated age comes to be larger due to this bias.

### 4.2. Estimation of the perceptual age with its labels

Results of the listening test gave us a unique value (label) of the perceptual age for each of the speakers, which was calculated as the averaged perceptual age over the subjects and could be rounded into an integer. If the bias problem is solved, this value can be used directly as $x$ in Equation 1. The bias problem was figured out by making approximation of $P(o|x)$ as follows.

$$P(o|x) \approx \max_i P(o|M_x^i), \tag{3}$$

where $M_x^i$ is a GMM of the $i$-th speaker who was estimated to be $x$ years old by the listening test. Although this approximation solved the bias problem, another problem remained intact. A range of age is limited from 0 to around 100. Therefore, if the maximum value of $P(o|x)$ is found in an extremely high range of age, the expected value tends to be smaller because there is almost no distribution over the maximum point on age axis. To solve this "limit" problem, we restricted the range of the expectation so that the operation was done only for $N$ values of age which showed the $N$ largest probabilities of $P(o|x)$. The case of $N$=1 corresponds to the widely-used maximum likelihood criterion. But $M_x^i$ is a speaker model and characterizes his/her age only roughly, and then, we considered that the expectation was necessary. Experimental verification of value of $N$ is described shortly and our consideration turns out to be correct.

### 4.3. Results and discussions

Results of the age estimation is shown in Figure 2. The perceptual age of an input speaker was estimated by using GMMs of all the other speakers. $P(o|x)$ in Equation 3 was obtained as normalized likelihood score, which is *not* in the logarithmic scale. The listening test was done with 30 subjects and 30 kinds of labels could be separately used for the estimation with another additional kind, which was obtained by averaging the 30 kinds of labels. Averaged correlation between the perceptual age defined by the listening and the automatically estimated age was 0.85 and the averaged labels (Figure 2) showed the largest correlation, which was 0.89. Value of $N$ was experimentally determined and $N$=11 was adopted in Figure 2, which indicates high necessity of the expectation. Dots of the estimated age are close to a $y$=$x$ line, but some distortions can be found in the figure. This is considered due to a fixed value of $N$ over the entire range of age. Adaptive control of $N$ should be required. Another approximation of $P(o|x)$ as $\overline{P(o|M_x^i)}$ over $i$ was tentatively examined but Equation 3 showed the better performance.

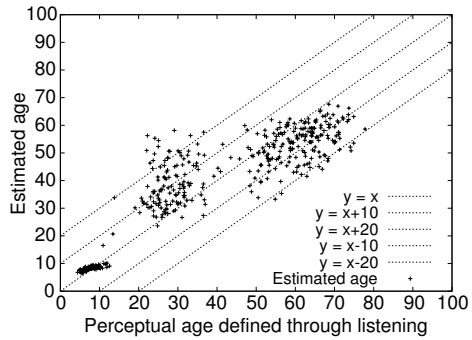


Figure 2: *Estimated age with labels of the perceptual age*

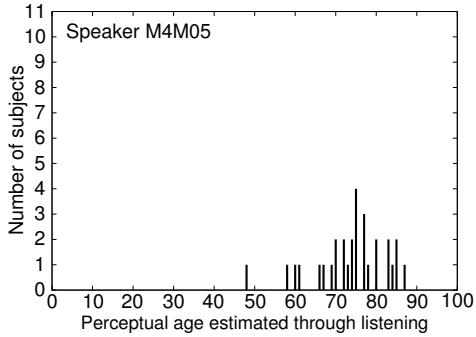


Figure 3: *A typical example of the age estimation by listening*

# 5. Estimation of the perceptual age with its distributions

## 5.1. Modeling of the perceptual age defined by the listening

In the previous section, the perceptual age was defined as a single value for each speaker. But results of the listening experiment showed the estimation naturally had some variances. In this section, the perceptual age is defined as a distribution and used for the estimation. Figure 5 shows a typical example of the human age estimation. From this figure, the distribution can be said to be properly modeled by a normal distribution.

## 5.2. Estimation of the perceptual age with its distributions

The following distribution $PA_0(x)$ was firstly estimated by using distributions of all the other speakers in the databases.

$$PA_0(x) = \frac{\sum_i P(o|M_i) g_i(x)}{\sum_i P(o|M_i)}, \tag{4}$$

where $g_i(x)$ is speaker $i$'s distribution function of the perceptual age, which is modeled by a normal distribution. Since the above equation directly uses $P(o|M_i)$, it comes to have a biased distribution due to the bias problem described in Section 4.1. When the perceptual age is given as label, a speaker group can be formed for each age so that the bias problem can be figured out in Equation 3. Speaker grouping is not impossible based on $g_i(x)$ but if the grouping is done, it means that $g_i(x)$ is viewed as discrete label. Based upon these considerations, we adopted Equation 4, which gives us the expected distribution over the speakers. The biased problem was solved in post-processing.

Equation 4 estimates the perceptual age distribution by referring to distance between the input speaker and the individual speakers in the databases. If the input speaker has the same distance to every speaker in the databases, he/she should be labeled as "completely unknown" about the age. Figure 4 shows

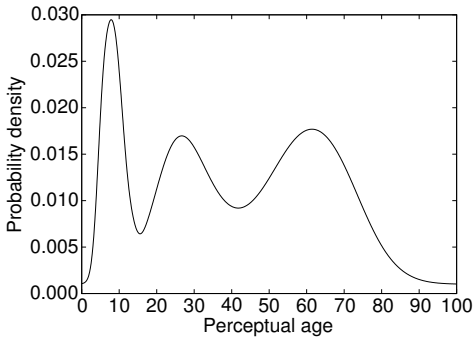


Figure 4: *Function to cancel the bias of the age distribution*

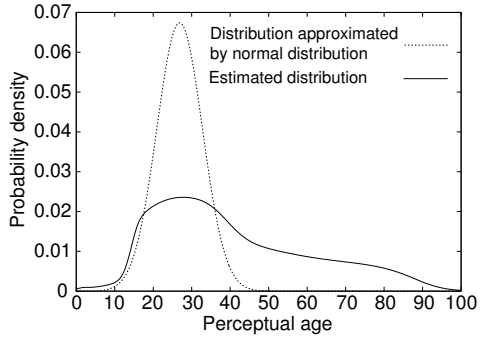


Figure 5: *A typical example of $PA(x)$ in JNAS database*

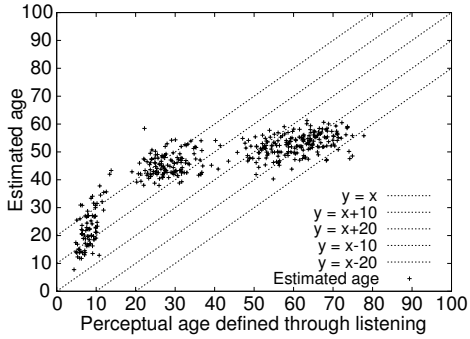


Figure 6: *Estimated age with distributions of the perceptual age*

$PA_0(x)$ with the same distance to every database speaker. The bias problem of Section 4.1 can be seen clearly in the figure. Furthermore, dependency of magnitude of variance on that of mean in the normal distributions of the perceptual age can also be found. By using this biased function ($BF(x)$) to cancel the inherent bias, $PA(x)$ was finally calculated as

$$PA(x) = \frac{PA_0(x)}{BF(x)}. \tag{5}$$

It should be noted that the minimum value of $BF(x)$ was artificially set to be 0.001 to avoid a zero-division problem.

Figure 5 shows a typical example of $PA(x)$ of an input speaker with his/her assumed age distribution approximated by a normal distribution in Section 5.1. Though $BF(x)$ solved the bias problem, the limit problem can be seen in the distribution. This speaker was judged to be about 27 years old by the 30 subjects and the upper range of age is wider than the lower range. If a single and representative value is estimated by doing an expectation operation on this distribution, this speaker comes to be older than 27. Figure 6 shows results of the perceptual age estimation, where x-axis represents the average of the human age

estimation and y-axis indicates the expected value of $PA(x)$ over the *entire* range of age. As is predicted above, speakers of the CHILDREN and JNAS databases tend to be estimated older and those of the SJNAS database tend to be estimated younger. In Section 4.2, this problem was solved by limiting the range for the expectation and the same solution was examined here.

### 5.3. Adaptive control of the range for the expectation

Extraction of a single and representative age value from a given distribution was carried out by expecting it on a limited range of the distribution. Firstly, a peak of the distribution was detected by a simple peak-picking method. If more than one peak were found, the peak with the largest probability density was selected. After that, for a range such that $x_p - x_c \leq x \leq x_p + x_c$, the expectation was done, where $x_p$ is age for the peak and $x_c$ is a control parameter to decide the range. As is described in Section 2.3, sharpness of the estimated age distribution depends upon its mean value. Then, in this study, $x_c$ was adaptively decided according to $x_p$. Figure 7 shows relation between mean and standard deviation of the human age estimation. The dotted line in the figure is a polynomial approximation curve of the relation, which is denoted as $f(x)$ here. Using $f(x)$, $x_c$ was calculated as $\alpha f(x_p)$, where $\alpha$ was determined experimentally.

### 5.4. Results and discussions

Figure 8 shows results of the age estimation by limiting the range for the expectation. Parameter $\alpha$ was determined so that correlation between the human judgment and the machine estimation was maximized. Correlation in Figure 8 is 0.88 and no improvement was found compared to that obtained in Figure 2, where the human judgment was used as discrete labels. However, it can be clearly found that dots of the estimated age are much closer to a $y{=}x$ line than those in Figure 2. On the other hand, we can see more dots located away from the line. Reasons for these data are discussed later. Figure 9 shows results of the age estimation only by using the peak value, namely, based on the maximum likelihood criterion. Correlation was 0.87 and almost no difference was found between Figures 8 and 9. This result implies that if the human judgment was modeled as a distribution and used for the perceptual age estimation, extraction of a single and representative age value for an input speaker can be done effectively based on the ML criterion.

As for reasons for dots distant from a $y{=}x$ line, we listened intensively to these data. But we, humans, could not find any difficulty to estimate the age for these speakers. In our previous study[5], we confronted the same kind of problem and we introduced prosodic features to characterize the agedness more adequately and some improvements were actually found. We consider that some of these speakers can be treated appropriately with prosodic features, which is one of the future works.

## 6. Conclusion

In this paper, a technique was proposed to estimate the perceptual age of an input speaker. Two methods were examined. The first method used results of the human age judgment by listening as discrete labels and the other used them as distributions. Both methods showed almost the same correlation between the human judgment and the machine estimation. The latter showed more graceful relation between the two but some speakers were found to be difficult to estimate the age only based upon speaker modeling techniques. As future works, we're planning to introduce speech rate and power perturbation, which proved to be

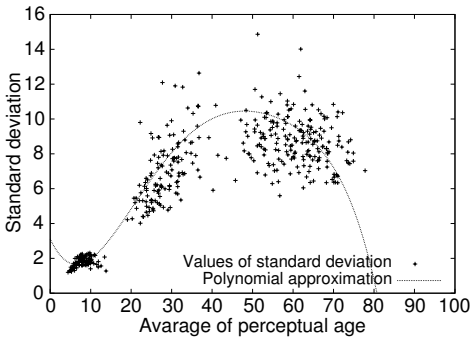


Figure 7: *Standard deviation of age as a function of its mean*

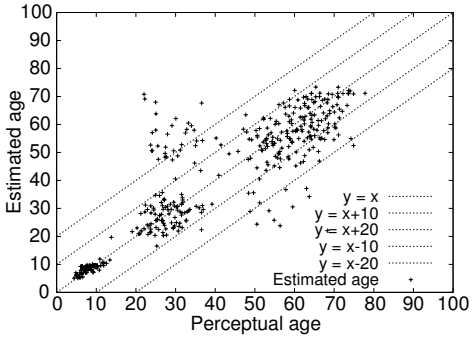


Figure 8: *Age estimation based on the expectation*

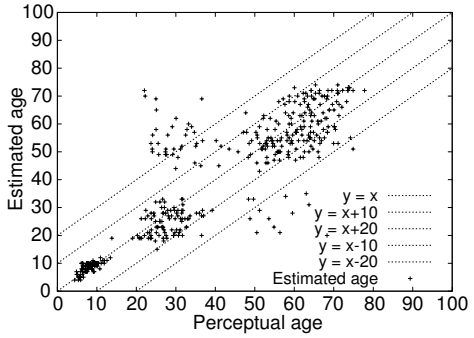


Figure 9: *Age estimation based on the ML criterion*

effective in our previous study, as additional parameters and integrate this technique to some existing systems to realize more meticulous control of their user-interface and dialogue strategy.

## 7. References

[1] M. Turk *et al*., "Perceptual User Interfaces," Communications of the ACM, vol.43, no.3, pp.33–34 (2000)

[2] A. Pentland, "Perceptual intelligence," Communications of the ACM, vol.43, no.3, pp.35–44 (2000)

[3] C. Müller *et al*., "Adapting multimodal dialog for the elderly," Proc. ABIS-Workshop on Personalization for the Mobile World (2002)

[4] T. Konuma *et al*., "A study of the elder speech recognition," Report of Fall Meet. Acoust. Soc. Jpn., 2-Q-1, pp.117–118 (1997)

[5] N. Minematsu *et al*., "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," Proc. ICASSP, pp.137–140 (2002)

[6] A. Baba *et al*., "Elderly acoustic model for large vocabulary continuous speech recognition," Proc. EUROSPEECH, pp.1657–1660 (2001)