Modeling the perception of children's age from speech acoustics

Santiago Barreda, and Peter F. Assmann

Citation: The Journal of the Acoustical Society of America **143**, EL361 (2018); doi: 10.1121/1.5037614
View online: https://doi.org/10.1121/1.5037614
View Table of Contents: http://asa.scitation.org/toc/jas/143/5
Published by the Acoustical Society of America

Articles you may be interested in

Visual perception of vowels from static and dynamic cues
The Journal of the Acoustical Society of America **143**, EL328 (2018); 10.1121/1.5036958

Musician effect on perception of spectro-temporally degraded speech, vocal emotion, and music in young adolescents
The Journal of the Acoustical Society of America **143**, EL311 (2018); 10.1121/1.5034489

Focus prosody of telephone numbers in Tokyo Japanese
The Journal of the Acoustical Society of America **143**, EL340 (2018); 10.1121/1.5037360

Speaking rhythmically improves speech recognition under “cocktail-party” conditions
The Journal of the Acoustical Society of America **143**, EL255 (2018); 10.1121/1.5030518

Understanding dysrhythmic speech: When rhythm does not matter and learning does not happen
The Journal of the Acoustical Society of America **143**, EL379 (2018); 10.1121/1.5037620

Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception
The Journal of the Acoustical Society of America **143**, EL372 (2018); 10.1121/1.5037615

CrossMark

Modeling the perception of children's age from speech acoustics

Santiago Barreda[a)]
Department of Linguistics, University of California, Davis, Davis, California 95616, USA
sbarreda@ucdavis.edu

Peter F. Assmann
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas 75080, USA
assmann@utdallas.edu

Abstract: Adult listeners were presented with /hVd/ syllables spoken by boys and girls ranging from 5 to 18 years of age. Half of the listeners were informed of the sex of the speaker; the other half were not. Results indicate that veridical age in children can be predicted accurately based on the acoustic characteristics of the talker's voice and that listener behavior is highly predictable on the basis of speech acoustics. Furthermore, listeners appear to incorporate assumptions about talker sex into their estimates of talker age, even when information about the talker's sex is not explicitly provided for them.

© 2018 Acoustical Society of America
[DDO]
Date Received: November 22, 2017 **Date Accepted:** April 25, 2018

1. Introduction

When attending to a recording of an unfamiliar voice, listeners form an immediate impression of the speaker's sex, age, and physical size, along with other personal attributes collectively referred to as indexical properties (Abercrombie, 1967). Several studies have found that listeners can estimate the age of the speaker with varying degrees of success depending on the speech material, the characteristics of the speech sample, and the task [for reviews see Linville (2001) and Schötz (2007)].

The perception of age in children's voices is particularly interesting because age-related changes in the voice throughout childhood are correlated with substantial changes in physical size (Kuczmarski *et al.*, 2002). Further, there can be a substantial variation in growth rates of individual components of the vocal tract (Vorperian *et al.*, 2009), resulting in a substantial variation in vocal tract geometries for children of a given age, and for a single child at different ages. Despite this, some general statements about the characteristics of children's voices can be made. First, children's vocal tracts are shorter than those of adults, leading to higher formant frequencies, and their larynges are smaller, resulting in a higher average fundamental frequency ($f0$). Furthermore, there are age-related decreases in $f0$ and formant frequencies, though these changes are larger in boys compared to girls. This leads to age-dependent variation in speech acoustics as a function of sex (Lee *et al.*, 1999; Perry *et al.*, 2001). Finally, vowel duration also shows systematic decreases with age in children, though this pattern is roughly the same for males and females (Lee *et al.*, 1999; Harnsberger *et al.*, 2006; Assmann *et al.*, 2013). These properties contribute to the perception of phonetic contrasts but are also linked to the age, sex, and size of the speaker.

Although there have been many studies investigating the correlations between age and speech acoustics in children, the literature on the perception of vocal age in children is relatively sparse, with most studies focused on the perceived age of adults rather than child talkers. Amir *et al.* (2012) investigated the perception of age in a sample of speech recorded from 120 children, including boys and girls from six age groups between 8 and 18 yrs. They found better than chance accuracy for age classification (perceived age within ± 2 yrs of chronological age), with more accurate responses for full sentences compared to isolated vowels. However, they did not relate perceptual judgments to acoustic properties.

The present study focuses on investigating the acoustic correlates of perceived age in children's voices. Given that age-related changes in $f0$ and formant frequencies

[a)]Author to whom correspondence should be addressed.

follow different trajectories in males and females, we examine whether knowledge of the sex of the talker influences listeners' judgments of their age.

2. Methods

2.1 Subjects

The listeners were 24 undergraduate students at the University of Texas at Dallas who received experimental credits for their participation. All were native speakers of American English with normal hearing as determined by pure-tone screening at 500, 750, 1000, 2000, and 4000 Hz. Half were informed about the sex of the speaker on each trial prior to responding; the other half were not.

2.2 Stimuli

The stimuli were 420 recorded syllables drawn from a children's speech database (Assmann *et al.*, 2008). Five boys and five girls were included at each age level between 5 and 18 yrs, for a total of 140 talkers. Each talker contributed three syllables: /hid/ ("heed"), /hɑd/ ("hod"), and /hud/ ("who'd").

2.3 Procedure

Stimuli were presented monaurally to the right ear via headphones at an average level of 68 dB sound pressure level (A-weighting) using Tucker-Davis System 3 and RP2.1 hardware (Tucker Davis Technologies, Inc., Alachua, FL). Each listener heard a different random sequence of the 420 stimuli. They adjusted a graphical slider on the computer screen to register their estimate of the speaker's age. Prior to the experiment, 24 familiarization trials (using stimuli similar to but distinct from those in the experiment) were presented with feedback. The experiment was self-paced and lasted about 50 min, with an optional break at the midpoint.

2.4 Description of models to be considered

A summary of the seven models being considered is presented in Table 1. All models predict age (either veridical or perceived) using three continuous predictors related to speech acoustics: talker geometric-mean formant frequency (GMFF) for the lowest three formants; the natural logarithm of the mean *f*0 averaged across the voiced portion of the syllable (G0); and average syllable duration. In addition, a subset of models contain a dummy variable indicating talker sex (0 = female, 1 = male), and the interaction between this predictor and the three acoustic predictors.

All formant measurements were derived from 50-ms frames sampled at the 20% point in the vowel. Acoustic measurements were averaged across all three syllables for each voice, as were age estimates provided by listeners. GMFF and G0 are expressed in natural log-transformed Hertz (log-Hz), while duration was expressed in tenths of seconds, in order to keep the time and frequency units of roughly equal magnitude.

Models predicting veridical age from acoustics (A, D) were fit using ordinary least-squares regression. Models predicting perceived age from acoustics (B, C, E, F) were fit using mixed-effects models using the lme4 package (Bates *et al.*, 2014) in R (R Core Team, 2016). All mixed-effects models included random effects for talker and listener, and random slopes for duration, G0, and GMFF. The significance of fixed-effect predictors was assessed using the lmerTest package (Kuznetsova *et al.*, 2015). Model predictions provided for the mixed-effects models (Table 2) were calculated using only fixed effects.

Table 1. Summary of models to be considered. Models incorporating talker-sex information included interactions between acoustic predictors and talker-sex.

		Talker-Sex Information		
Model	Dependent Variable	Available to Listeners	Available to Model	Acoustic Predictors
A	Veridical Age	N/A	No	Duration, G0, GMFF
B	Perceived Age	Yes	No	Duration, G0, GMFF
C	Perceived Age	No	No	Duration, G0, GMFF
D	Veridical Age	N/A	Yes	Duration, G0, GMFF
E	Perceived Age	Yes	Yes	Duration, G0, GMFF
F	Perceived Age	No	Yes	Duration, G0, GMFF

Table 2. Coefficient estimates for the different models outlined in Table 1. Models A and D predict veridical age, the remaining models predict perceived age. Asterisks indicate significant predictors at a $p < 0.05$ threshold. Residual MAE is the dependent variable for each model. Predictive MAE refers to the error in predicting veridical age using each model, and Listener MAE indicates average age estimation error across all talkers and listeners in each condition.

	Predictors									
Model	Dur	G0	GMFF	Male	Dur × Male	G0 × Male	GMFF × Male	Residual MAE	Predictive MAE	Listener MAE
(A)	**−1.65***	1.82	**−33.5***	—	—	—	—	2.03	—	—
(B)	**−1.19***	−0.70	**−24.8***	—	—	—	—	1.6	2.07	1.79
(C)	**−1.20***	−0.72	**−26.2***	—	—	—	—	1.43	2.04	1.8
(D)	**−1.10***	−1.58	**−35.2***	**1.59***	−0.21	0.72	**−9.81***	1.68	—	—
(E)	**−0.73***	**−4.06***	**−26.1***	**1.38***	−0.13	−0.51	**−6.75***	1.27	1.8	1.79
(F)	**−0.91***	**−2.92***	**−27.1***	**0.84***	0.01	−0.39	**−5.11***	1.3	1.82	1.8

3. Results

3.1 Relationship between veridical age and speech acoustics

As seen in Fig. 1, there were significant negative correlations between veridical age and G0 ($r = -0.62$, $p < 0.001$, $N = 140$), GMFF ($r = -0.75$, $p < 0.001$, $N = 140$), and duration ($r = -0.42$, $p < 0.001$, $N = 140$). There was a strong positive correlation between G0 and GMFF ($r = 0.86$, $p < 0.001$, $N = 140$), and moderate positive correlations between G0 and duration ($r = 0.32$, $p < 0.001$, $N = 140$), and GMFF and duration ($r = 0.30$, $p < 0.001$, $N = 140$).

3.2 Modeling veridical age from speech acoustics

The first column in Fig. 2 compares age predictions made by the acoustic models trained on veridical talker age. The coefficients for these models (A, D, Table 2) represent the optimal use of acoustic cues for estimating the ages of the talkers in the experiment, given the data and the structure of the model. Duration and GMFF were significant predictors of age in both of these models, while neither model has a significant effect for G0.

Model D, which includes talker-sex information, is substantially more accurate in predicting veridical age than model A, which does not (1.68 vs 2.03 residual mean absolute error, MAE). The primary difference between models A and D is in the main effect for talker sex and the sex × GMFF interaction. As can be seen in Fig. 1(a), males have a lower GMFF at every age, and a larger change in GMFF as a function of age. The sex-related terms in model D allow age to relate to GMFF in a sex-dependent manner. This results in a more accurate age estimation for both sexes, indicating that optimal prediction of children's age from speech acoustics will likely involve the consideration of talker sex.

3.3 Modeling perceived age from speech acoustics when talker sex is known

The second column in Fig. 2 compares age estimates when talker sex was known to listeners (models B, E), to the predictions made by models trained on these estimates. Three results indicate that listeners are using talker-sex information when

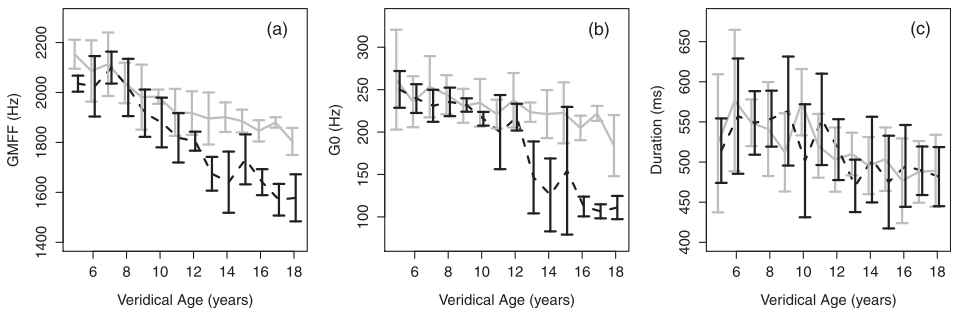

Fig. 1. (a) Mean GMFF as a function of age for males (dashed line) and females (solid line); (b) geometric-mean fundamental frequency (G0) as a function of age for males (dashed line) and females (solid line); (c) mean duration as a function of age for males (dashed line) and females (solid line). Error bars indicate one standard deviation across talkers. There were five talkers at each age for each sex.

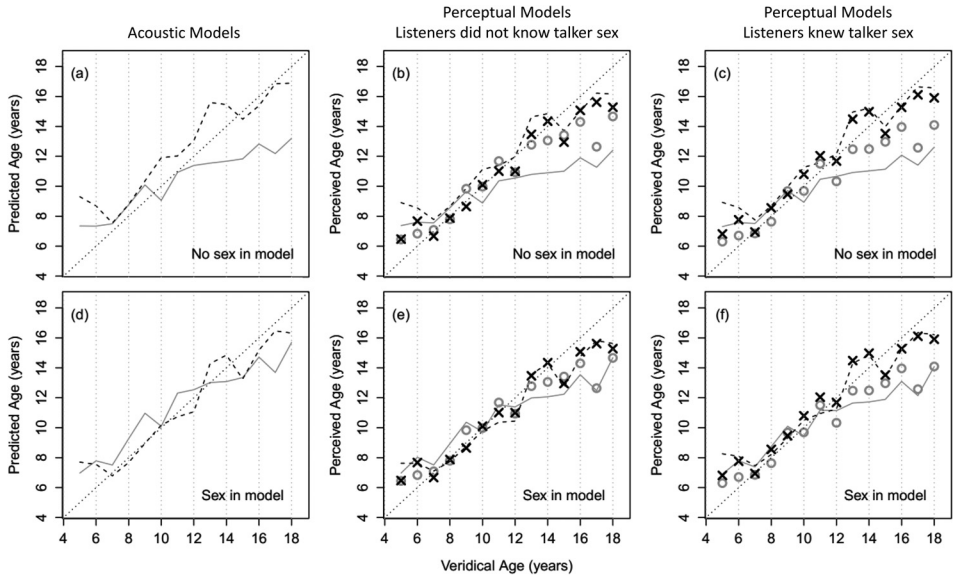

Fig. 2. Model predictions of veridical [(a) and (d)] and perceived [(b), (c), (e), and (f)] talker age by each of the models outlined in Table 1. Panel letters correspond to model letters in Table 1. Thin lines indicate model predictions (solid for females, broken for males), points indicate observed age estimates (circles for females, crosses for males). The dotted diagonal line indicates where predictions would match veridical talker age exactly.

estimating talker age. First, model E includes a significant effect for talker sex, and a significant talker sex × GMFF interaction. Second, the residual MAE for the prediction of perceived age is substantially smaller for the model that includes talker sex (1.27, model E) as compared to the model that does not (1.60, model B). Third, we may consider predictive MAE, which is accuracy in the prediction of veridical age using a model trained on perceived age. This measure essentially treats models as artificial listeners that estimate talker age based on speech acoustics. As seen in Table 2, the predictive MAE for model B (2.07) is larger than the MAE for the group of listeners who were given talker-sex information (1.79). However, the predictive MAE of model E is quite close (1.80). Taken together, the above offers strong evidence that listeners use talker-sex information when estimating the age of children from speech acoustics.

The coefficient pattern for model E is broadly similar to the optimal use of these cues for the dataset (model D), save for two noteworthy differences. First, the effects for GMFF are weaker. This may be due to the fact that GMFF is related to information related to talker vocal-tract length, which must be estimated by the listener and is not directly present in the acoustic signal. Second, the optimal model trained on veridical age (D) has no role for G0, while listeners do appear to use this cue.

3.4 Modeling perceived age from speech acoustics when talker sex is not known

The third column in Fig. 2 compares age estimates when talker sex was not known to listeners (models C, F), to the predictions made by models trained on these estimates. Although listeners were not explicitly given talker-sex information, the pattern of results suggest that listeners may have been using perceived talker sex (e.g., as a latent variable) when assessing the talker's age.

First, the pattern of coefficient values and significant effects is quite similar for models E and F, including significant effects for talker sex and a significant talker sex × GMFF interaction. Furthermore, the model including talker-sex information (F) has a smaller residual MAE (1.30) than the model that did not include this information (C, 1.43). Finally, the predictive MAE of the model including talker-sex information (F, 1.82) was more similar to the listener MAE (1.80) than the model that did not include talker-sex information (C, 2.04). In general, the model including talker-sex information did a better job of explaining and replicating listener behavior than the model with no sex information, even though listeners were not given explicit information about talker sex.

3.5 Errors in age estimates

In order to investigate general trends in error patterns, MAEs were calculated for each listener, for each combination of talker age and sex. A mixed-design analysis of

variance was carried out on these errors with talker sex and age as within-subject factors, and talker-sex information as a between-subject factor. Results show significant main effects for talker sex [$F(1,22) = 24$, $p < 0.001$] and age [$F(13\,286) = 24$, $p < 0.001$], but not for talker-sex information [$F(1,22) = 0.01$, $p < 0.001$]. Additionally, the age $\times$talker sex [$F(13\,286) = 57$, $p < 0.001$] and age $\times$ talker sex$\times$ talker-sex information interactions [$F(13\,286) = 4.4$, $p < 0.001$] were both significant.

To assess the relative importance of these effects on age-estimation errors, a regression model was fit to this data, pooled across all listeners. This model sought to predict the MAE in age estimation for each subject as a function of three categorical predictors and their interactions: talker age, talker sex, and talker-sex information. A type III analysis of variance carried out on this model indicates that these factors explain 55% of the variance in these errors. However, a large majority of this is explained by talker age (30%) and the talker sex $\times$ talker age interaction (21%), whereas the main effect for talker sex (2.8%) and the talker sex $\times$ information $\times$ talker age interaction (1.6%) explain only a small amount of variance in error patterns. This indicates that error patterns are dominated by talker age, with sex-specific differences in this pattern, and that explicit information about talker sex did not make listeners more accurate.

4. Discussion

Results demonstrate that adult listeners are able to estimate age from children's voices fairly accurately (MAE = 1.8 yrs) when presented with isolated syllables, and that listener behavior is highly predictable on the basis of speech acoustics. GMFF provides a more reliable basis for age judgments than G0, while duration makes a relatively small but significant contribution. Results also indicate that listeners are likely using talker sex information in their judgments, even when this information is not explicitly provided. This suggests that listeners are estimating talker sex from speech and using this information to arrive at more accurate age judgments.

Previous reports indicate that listeners can estimate children's sex from speech at a higher than chance level: 65% correct identification for children as young as 4 yrs old increasing to more than 90% correct for post-pubescent children (Perry *et al.*, 2001; Amir *et al.*, 2012). Perry *et al.* (2001) suggest that this is primarily on the basis of formant frequencies since males have lower formant frequencies than females for any given age group, whereas *f*0 does not allow talker sex to be discriminated until after puberty (as in the data presented in Fig. 1). However, for many pre-pubescent male age levels, there is a group of older females with broadly similar speech acoustics. Given this, the reasonably accurate gender estimation for children reported in Perry *et al.* (2001) and Amir *et al.* (2012) suggests that listeners may benefit by taking into account talker age when estimating the sex of the child.

For example, in our data the mean GMFF for 10-yr old males is quite similar to that for 15-yr old females (7.54 log Hz, 1881 Hz for both), as is their mean G0 (5.37 and 5.41, respectively, 214 and 224 Hz) meaning sex and age are expected to be ambiguous for speakers in these groups. However, the correct age and sex decisions will tend to be correlated: a talker with these speech acoustics could be a 10-yr old male or a 15-yr old female, but not a 15-yr old male or a 10-yr old female. In this way, when talker sex and age decisions are made jointly, errors can be reduced by considering the covariance between talker categories with respect to speech acoustics. As a result, the perception of sex and age from children's speech can both benefit from being jointly estimated from acoustics, and the results presented here suggest that listeners are engaging in just this sort of behavior.

Additional cues related to the glottal source (e.g., breathiness) which vary with age and sex (Iseli *et al.*, 2007), may contribute to listener judgments of age directly, or indirectly by informing sex judgments. Preliminary modeling results did not reveal significant contributions for such measures in the data set reported here. However, further research is needed investigating the perception of age and sex from children's voices, and the ways that judgments of these talker characteristics may influence each other, and we are currently pursuing these topics in further experiments.

Acknowledgments

This research was supported in part by a grant from the National Science Foundation (Grant No. 1124479, P.F.A.). We would like to thank Terry Nearey for his helpful comments and Daniel Hubbard for assistance in data collection.

References and links

Abercrombie D. (**1967**). *Elements of General Phonetics* (Edinburgh University Press, Edinburgh).

Amir, O., Engel, M., Shabtai, E., and Amir, N. (**2012**). "Identification of children's gender and age by listeners," J. Voice **26**(3), 313–321.

Assmann, P. F., Nearey T. M., and Bharadwaj, S. (**2008**). "Analysis and classification of a vowel database," Canadian Acoust. **36**(3), 148–149.

Assmann, P. F., Nearey, T. M., and Bharadwaj, S. V. (**2013**). "Developmental patterns in children's speech: Time-varying spectral change in vowels," in *Vowel Inherent Spectral Change*, edited by G. S. Morrison and P. F. Assmann (Springer-Verlag, Heidelberg), pp. 199–230.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2014**). lme4: Linear mixed-effects models using Eigen and S4, R package version 1.1-12.

Harnsberger, J. D., Shrivastav, R., Brown, W. S., Jr., Rothman, H., and Hollien, H. (**2006**). "Speaking rate and fundamental frequency as speech cues to perceived age," J. Voice **22**, 58–69.

Iseli, M., Shue, Y.-L., and Alwan, A. (**2007**). "Age, sex, and vowel dependencies of acoustic measures related to the voice source," J. Acoust. Soc. Am. **121**, 2283–2295.

Kuczmarski, R. J., Ogden, C. L., Guo, S. S., Grummer-Strawn, L. M., Flegal, K. M., Mei, Z., Wei, R., Curtin, L. R., Roche, A. F., and Johnson, C. L. (**2002**). "2000 CDC growth charts for the United States: Methods and development," Vital Health Stat. **11**, 1–190.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2015**). Package 'lmerTest'. R package version, 2.0-30.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468.

Linville, S. E. (**2001**). *Vocal Aging* (Singular Thomson Learning, San Diego, CA).

Perry, T. L., Ohde, R. N., and Ashmead, D. H. (**2001**). "The acoustic bases for gender identification from children's voices," J. Acoust. Soc. Am. **109**, 2988–2998.

R Core Team (**2016**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ (Last viewed January 10, 2018).

Schötz, S. (**2007**). "Acoustic analysis of adult speaker age," in *Speaker Classification I*, edited by C. Müller (Springer-Verlag, Berlin, Heidelberg), LNAI 4343, pp. 88–107.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., Ziegert, A. J., and Gentry, L. R. (**2009**). "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study," J. Acoust. Soc. Am. **125**, 1666–1678.