Modeling perceived vocal age in American English

James D. Harnsberger,1 Rahul Shrivastav,2 and W.S. Brown, Jr.2

1 Program in Linguistics, University of Florida, USA
2 Department of Speech, Language, and Hearing Sciences, University of Florida, USA
jharns@ufl.edu, rahul@ufl.edu, wsbrown@ufl.edu

Abstract

An acoustic analysis of voice, articulatory, and prosodic cues to perceived age was completed for a speech database of 150 American English speakers. Perceived ages were submitted to multiple linear regression analyses with measures of acoustic correlates of: voice quality, articulation, fundamental frequency, and prosody. The fit between predicted and actual perceived ages from the resulting models varied by speech material and gender, with female vocal ages being the easiest to predict. Articulation, pitch, and speaking rate measures were the most predictive in female voices, while, for male voices, the observed ranking was: speaking rate, voice quality, and pitch.

Index Terms: Paralinguistic and nonlinguistic cues, Human speech and sound perception, Prosody

1. Introduction

Vocal aging refers broadly to the changes in all major systems of speech production, which can be defined broadly as voice, articulation, and prosody. Voice here refers to the control of vocal fold vibration in speech production. Problems with this system due to old age can be seen in 1) a lack of control of the rate of vibration [1]-[3], 2) systematic shifts in the rate – varying by gender [1], [4]-[6], and 3) abnormalities in the adduction of vocal folds or in the folds themselves, resulting in excessively breathy, hoarse, strained, creaky, or tremorous voices [7]-[8]. The term articulation refers to the control of the articulators within the supralaryngeal tract. Articulatory effects of aging may incorporate the acoustic consequences of laryngeal lowering with age, vowel centralization, higher misarticulation and substitution rates, deletions, insertions, and so on [9]-[12]. Finally, the term prosody here includes the control and coordination of voice and articulation over longer periods of time. With age, aspects of prosody can change, such the rate of speaking, the frequency of occurrence of pauses or hesitations, and the overall intensity of utterances [12]-[14].

While these acoustic studies of phonation and articulation have been proved valuable in the study of vocal aging, to date only a few models have been developed for American English that attempt to predict the perceived age of a speaker. That is, individual studies have examined different subsets of acoustic cues, but it has not been as common to integrate a larger set of data. For those that have, studies have typically relied either on perceptual judgments of acoustic correlates of perceived age [15]-[16], actual acoustic measures [2], or a mix of both [17], and have had limited success. More recent efforts at modeling vocal age have involved languages other than American English, including Swedish [18], German [19]-[20], and Japanese [21]. Minematsu et al. [21] examined, separately, spectral information (in the form of Mel-frequency cepstral coefficients (MFCC) and power), speaking rate (in morae and in rate of peaks in delta-MFCC vectors), and in an unspecified measure of fundamental frequency perturbation in order to classify correctly voices judged to been subjectively elderly. A good identification rate (91%) was achieved using only spectral information. However, no more specific cues were examined, and the relative importance of competing cues was not the topic of the study. A wider range of measures were employed in Brückl’s [19] modeling effort for in German, although not all of these were summarized to permit direct comparisons of cues. Brückl [19] collected longitudinal data across a five year span for a set of German female speakers (nine total) varying in age from 26 – 87, recording both read and spontaneous speech. Statistically significant findings are only reported for speaking rate measures and formants (F1 – F3, with F1 lowered and F2 and F3 raised), although other trends are alluded to. More direct comparisons of cues could be made by Schötz [18], who modeled chronological age in a database of 428 Swedish voices (gender unspecified) producing single words, and was quite comprehensive in the range of acoustic measures taken. Out of a 51 item feature set, 21 features showed significant, moderate correlations with chronological age, ranging between r = 0.40 to 0.52. Of these, spectral cues including the third and fourth formants were most predictive, while duration and voice quality measures did not play a significant role in estimating Swedish ages. One limitation of Schötz’s [18] approach, however, was the use of word-level data.

Both the success and limitations of earlier modeling efforts in these languages provide guidance for the development of new, more robust models of perceived age. First, any modeling effort would benefit from a range of speech samples in duration, particularly given the frequently observed differences in the fit between chronological and perceived age. Second, an ideal speech database would be balanced by gender to better examine fundamental frequency effects, and incorporating a large number of voices to maximize generalizability. Finally, given that so many speech cues have been observed to change with age (pitch, voice quality, speaking rate, vowel formants), a reasonably large set of acoustic measures should be incorporated within the same modeling effort. Therefore, the purpose of the present study was to apply this modeling process to American English.

2. Methods

A variety of speech samples were produced by 50 chronologically young (18 – 30 years), 50 middle aged (40 – 55 years), and 50 older talkers (62 -92 years). The mean ages for each group were 21, 48, and 79, respectively. Within each set of 50 talkers, half were male and half were female. All were recruited from the Gainesville, Florida, USA area. No attempt was made to control for dialect background, although all three groups were predominately represented by individuals who had lived in many regions of the country over their lifespans. All participants were native speakers of English with no known history of speech or hearing problems.

The speech materials included the Grandfather Passage, the Rainbow Passage, sixteen sentences taken or adapted from
the Speech Perception in Noise test, three sustained vowels ([a], [i], and [u]), and two diadodes. Sustained vowels were deemed most useful for testing voice quality measures (such as jitter or shimmer). Diadodes represented both a word-level stimulus as well as one that both a) taxed the phonatory and articulatory control systems of older adults and b) induced a hypoarticulated speaking style, in contrast with the hyperarticulated sustained vowels and citation-style sentences and passages. Sentences were incorporated to represent longer speech samples that incorporate prosodic variation while being brief enough to use in perceptual testing. Finally, passages were recorded representing a long, less monitored speaking style. Talkers were recorded in a quiet environment using a head worn microphone fixed at a constant distance from the corner of the mouth. Of all of these samples, the sixteen SPIN sentences were submitted to 147 listeners for age estimation for use in modeling.

3. Acoustic Analysis and Modeling

3.1. Features

Thirty three acoustic measures of pitch, voice quality, articulation, and prosody cues were taken, varying somewhat by speech material type. These cues were selected to represent a full range of phonetic measures (i.e., ones derived directly from articulatory or aerodynamic properties of speech) that have been examined individually or in smaller sets in previous studies, but not reported together for a single speech database. Fundamental frequency measures included mean, median, and standard deviation, using up to three different algorithms. Voice quality was assessed using four jitter measures, four shimmer measures, three metrics of noise (harmonics-to-noise ratio (two algorithms) a breathiness algorithm), cepstral peak prominence with smoothing (CPP-S), and six measures of degree or proportion of periodicity: active, autocorrelation, unvoiced (two algorithms), voice breaks (count and degree), and voiced. Three prosody cues were taken automatically: the proportion of silence in the signal, the overall duration of the sample, and the standard deviation of the rmss intensity measure, calibrated by the average intensity of the speech sample. It should be noted that not all voice measures were taken from each category of speech material (e.g., jitter and shimmer measures were only taken from the sustained vowels).

Articulatory effects of aging were assessed by measuring the first three formant frequencies of the three sustained vowels ([a], [i], and [u]) and corresponding target vowels in the diadodes and a subset of the SPIN sentences. Formant measures were taken from a combined FFT and LPC (14-coefficient autoregressive analysis) display, with a 25 ms analysis window at the temporal midpoint of the vowel. These frequencies in Hertz were converted to Barks and then used in two types of calculations: formant distance and vowel space size. Formant distance was developed to quantify any age group differences in relative location of the vowel space in order to observe either overall formant lowering or any other systematic differences. It was calculated as follows: first, a reference point was established based on the mean F1 – F3 (in Barks) across all vowels produced by all 150 speakers. Second, difference scores were calculated for each formant, for each vowel, produced by each speaker relative to the reference point. These difference scores were averaged across all three vowels of a speaker for a given formant. Thus, for each speaker, an F1, F2, and F3 distance score was generated. For estimating vowel space size for a given speaker, distances in an F1-F2 Bark space were calculated for all three pairings of all vowels. These three distances were then averaged for a final score for mean vowel space size.

3.2. Modeling

The results of the acoustic analysis are reported here in Table 1 not with descriptive information for each measure, but rather in terms of the Pearson product-moment correlation coefficients between each cue and the perceived age of the speakers. The table lists, for each type of speech material, all significant correlation coefficients. Any cells left blank represent nonsignificant correlations, while shaded cells represent cues not measured for a given speech material type (typically the passages).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 mean (algorithm)</td>
<td>-0.49**</td>
<td>-0.47**</td>
<td>-0.45**</td>
<td>-0.50**</td>
<td>-0.39**</td>
<td>-0.35**</td>
<td>-0.49**</td>
<td>-0.44**</td>
</tr>
<tr>
<td>F0 (Distance)</td>
<td>0.55**</td>
<td>0.49**</td>
<td>0.49**</td>
<td>0.44**</td>
<td>0.51**</td>
<td>0.45**</td>
<td>-0.24**</td>
<td>-0.21**</td>
</tr>
<tr>
<td>F0 (Distance)</td>
<td>0.42**</td>
<td>0.27**</td>
<td>0.38**</td>
<td>0.35**</td>
<td>0.32**</td>
<td>0.30**</td>
<td>-0.23**</td>
<td>-0.19**</td>
</tr>
<tr>
<td>Overall Duration</td>
<td>0.23**</td>
<td>-0.24**</td>
<td>0.42**</td>
<td>0.44**</td>
<td>0.36**</td>
<td>0.62**</td>
<td>0.38**</td>
<td>0.76**</td>
</tr>
<tr>
<td>Silence (Proportion)</td>
<td>0.36**</td>
<td>-0.25**</td>
<td>0.36**</td>
<td>0.35**</td>
<td>-0.29**</td>
<td>-0.23**</td>
<td>0.29**</td>
<td>0.33**</td>
</tr>
<tr>
<td>Active (Proportion)</td>
<td>0.21**</td>
<td>-0.25**</td>
<td>0.46**</td>
<td>0.35**</td>
<td>-0.24**</td>
<td>-0.29**</td>
<td>0.29**</td>
<td>0.33**</td>
</tr>
<tr>
<td>Shimmer (Mean)</td>
<td>0.53**</td>
<td>0.45**</td>
<td>0.45**</td>
<td>0.48**</td>
<td>-0.20**</td>
<td>-0.39**</td>
<td>-0.45**</td>
<td>-0.40**</td>
</tr>
<tr>
<td>CPP-S</td>
<td>0.31**</td>
<td>0.21**</td>
<td>0.37**</td>
<td>0.32**</td>
<td>-0.29**</td>
<td>-0.20**</td>
<td>0.29**</td>
<td>0.33**</td>
</tr>
<tr>
<td>Unvoiced (Proportion)</td>
<td>0.32**</td>
<td>0.28**</td>
<td>0.32**</td>
<td>0.30**</td>
<td>-0.30**</td>
<td>-0.29**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voice Breaks (Count)</td>
<td>0.53**</td>
<td>0.36**</td>
<td>0.27**</td>
<td>0.25**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voice Breaks (Durations)</td>
<td>0.23**</td>
<td>-0.41**</td>
<td>0.25**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocal Duration</td>
<td>0.25**</td>
<td>-0.34**</td>
<td>0.52**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1 (Distance)</td>
<td>0.58**</td>
<td>-0.21**</td>
<td>-0.20**</td>
<td>-0.29**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2 (Distance)</td>
<td>0.43**</td>
<td>-0.36**</td>
<td>-0.30**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vowel Space (Mean)</td>
<td>0.37**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Turning first to the female voices, most but not all pitch, prosodic, voice quality, and articulation cues showed modest (±0.25 to ±0.69), significant correlations in the predicted direction. Mean and median fundamental frequency significantly lowered with age, while F0 variation increased in most measures with the sustained vowels and diadodes in particular. Voice quality cues were also relevant to perceived age: jitter and shimmer (however calculated) increased with perceived age in sustained vowels; CPP-S correlated with age in diadodes, sentences, and passages; breathiness and HNR were also significant, but not across all materials. Measures related to automatic assessment of periodicity were less consistent and more weakly correlated with perceived age than the other voice quality measures, with the possible exception of autocorrelation. Speaking rate (listed here as Overall Duration) correlated strongly with perceived age in all materials except for the sustained vowels, in which this measure is less interesting since it merely reflects speakers’ capability of following instructions to produce a vocalization of a fixed duration (five seconds in this study).
It should also be noted that some classes of cues, while significantly correlated with perceived age, nevertheless did not follow the predicted patterns shown in prior studies. Intensity variation did not consistently increase with age across all speech materials examined. The four articulation measures also fall into this category. Recall that laryngeal lowering in the elderly is hypothesized to be both common enough and significant enough to observe formant lowering in the articulation of vowels and consonants. In this database, a significant lowering trend was observed in F1 and F2 across all speech material types, but not F3, which actually displayed a very modest opposite trend in the tokens sampled from the sentence material. Finally, perceived age did not correlate with mean vowel space size with these female voices.

Male voices are reported separately from female voices not only because of expected differences in mean F0 measures, but also in the larger trends observed across classes of cues. In general, fewer measures correlated significantly with perceived age in males than females, and particularly, articulatory information was less relevant. Several interesting trends were observed. First, mean and median fundamental frequency significantly increased with age, as predicted, in sustained vowels and diadodes, but less obviously for longer speech samples. F0 standard deviation did not show consistent trends across materials and measures. Second, speaking rate showed some of the strongest, significant correlations in sentence and passage material. Third, for the voice quality and voice periodicity measures, CPP-S proved to be most predictive of perceived age; all other measures had either lower correlation coefficients or were not significant across all speech material types. Finally, among the articulation measures, vowel space size trends were not significant, with the exception of hypoarticulated vowels in the diadodes.

The trends reported above are of interest in and of themselves, but it is important for the development of a formal model of vocal aging to predict perceived age and to estimate residuals. The degree to which a model fails to capture variance in the perceived age data will demonstrate how much more ambitious we must be in future studies in selecting candidate acoustic correlates. This modeling was approached in two ways. First, a subset of the “best” measures was developed, with only one measure representing a cue of interest, and these measures were submitted to a multiple linear regression with perceived age. The list of best measures included: F0 mean and σ, Overall Duration, RMS Intensity σ, Breathiness, CPPS, Jitter (Local), Shimmer (local, dB), F1, F2, F3, Vowel Space (Mean). A measure would be selected as the “best” representative on the basis of the strength of its correlation across multiple speech samples types. This analysis permitted a more direct comparison of the relative importance of individual variables in the analysis by eliminating measures that are obviously correlated with one another (e.g., mean and median fundamental frequency). Second, all significant measures from Table 1 were submitted in a single multiple linear regression model (for each gender), including such intercorrelated measures as the three F0 means and F0 median, the F0 standard deviation measures, and so on. The use of multiple measures for the same cue afforded the opportunity to potentially reduce the residuals associated with each of those measures independently, at the cost of accuracy in comparing the relative importance of individual cues due to multicollinearity. Figure 1 provides the Pearson r coefficients for the best measures while Figure 2 gives the corresponding coefficients for all of the significant measures from Table 1. All correlation coefficients were significant and represent the extent to which these sets of phonetic measures accounted for the actual perceived ages.

Several observations can be made from these analyses. Turning first to the best measures analysis, the coefficients elicited for the female voices ranged from r = 0.77 to 0.87, with longer speech samples showing the best fits. Longer speech samples also resulted in better model performance for the male voices, with r values increasing from 0.62 to 0.88 from sustained vowels to full passages, even though the passage database did not include many of the voice quality measures (although articulatory measures were included, namely, those taken in the comparable SPIN sentence database). The addition of extra, intercorrelated measures, with the results shown in Figure 2, appeared to improve the performance of the sustained vowel models, with r values of 0.73 (males) and 0.92 (females). However, gains were more limited at longer speech sample types. Overall, modeling was successful using the cues measured, across a variety of read speech materials and with both genders, although significant room for improvement certainly remains.

### 4. Discussion

This study represents an effort to model perceived age in American English through the use of phonetically-grounded measures of pitch, voice quality, articulation, and prosody. Across speech materials and analyses, female perceived ages were easier to predict than male perceived ages, and...
correlation coefficients only improved modestly with longer speech samples. The strength of the correlations approached, but fell a little short of, a threshold established by Ryan and Burk [17] of 0.96 for American English, whose model relied on perceptual judgments rather than acoustic measures for some of its most important variables, particularly voice quality and consonant misarticulation.

The results of this study can also been compared to past work in terms of the relative importance of classes of cues. In this study, the relative importance of different cues varied substantially. Nevertheless, some trends can be cited. Generally speaking, for both genders, correlations were stronger in longer speech samples. For female voices, vowel articulation, pitch, and speaking rate cues showed comparable correlations with perceived age, followed by voice quality as represented by the breathiness measure. Other voice quality measures, such as the jitter, shimmer, periodicity measures; CPP-S, and mean vowel space, were more weakly correlated and/or inconsistently so across the different types of speech samples. This ordering varies substantially from that observed by Linville and Fisher [2], and it likely reflects the inclusion of multiple variables in the analysis and the use of longer speech samples. This ordering shares some similarity to Schötz’s [18] ranking of formant information (particularly F3 and F4) over pitch (fundamental frequency). However, voice quality and speaking rate still played a significant role in the estimation of female ages. There are many reasons why these rank orderings may differ, apart from the obvious language difference. For example, this study treated male and female voices separately. Schötz’s [18] model for Swedish pooled across gender and relied on word-level stimuli, which undoubtedly reduced the role speaking rate, for instance, played in age estimation.

For male voices, the strongest correlations were observed with speaking rate in the longer speech samples; voice quality (represented by CPP-S) and mean/median fundamental frequency followed in importance. Duration information also appeared to be more predictive in male voices relative to pitch, a pattern that did not carry over to the female voices. The gender asymmetry, however, is something of a surprise. It is at odds, for instance, with Winkler’s [20] observation that variations in speaking rate in German voices have a much more dramatic impact on perceived age than variations in fundamental frequency across both genders. That may indicate language- or culture-specific differences that remain uninvestigated. Finally, for male voices, articulatory differences due to age were weak in comparison to female voices, at least as measured in this study. Indeed, the articulatory effects observed in both genders beg for explanation. It has hypothesized that age-related differences in vowel formants may instead have a sociocultural origin [2]. That is, generational dialects exist within American English that may, in turn, serve as cues to perceived age by listeners. The fact that vowel formants were so predictive in this study of female voices particularly has a parallel in studies of dialect variation, in which younger female speakers commonly are innovators when it comes to sound changes. If there are age and gender effects in the adoption of new phonological forms, then it is reasonable to expect that listeners may exploit this in age estimation.

5. Conclusions

This study concerned an attempt to model perceived age using common acoustic correlates to articulation, pitch, voice quality, and prosodic effects of the aging process. Model performance varied by speech material and gender, with female vocal ages being the easiest to predict ($r = 0.81$ to $0.92$) as compared with males ($r = 0.73$ to $0.88$). Improvement to the models was suggested by further investigation in articulatory effects of aging process.

6. Acknowledgements

The following individuals should be acknowledged for their assistance in data collection and processing: Dot Bourgoise, Shelley Hicks, Anna Rowe, and Christine White.

7. References