2pSCb25. Quality of older voices processed by hearing aids: Acoustic factors explaining inter-talker differences

Huiwen Goy*, M. K. Pichora-Fuller, Pascal Van Lieshout and Kathryn Arehart

*Corresponding author's address: Department of Psychology, University of Toronto, 3359 Mississauga Road North, Mississauga, L5L 1C6, Ontario, Canada, huiwen.goy@utoronto.ca

Hearing aid signal processing algorithms are often evaluated with professional recordings of voices. However, hearing aid users often listen to older talkers who may have poorer voices than younger talkers. The purpose of this study was to quantify the extent to which the acoustic and perceptual consequences of hearing aid digital signal processing algorithms differ for talkers that vary in their vocal characteristics. There were five talkers (2 male and 3 female) selected from a larger database of 60 talkers; their voices had good, moderate or poor quality based on perceptual data from younger and older listeners. The voices were presented in quiet and in the presence of babble noise (-5 to +20 dB SNR). The voices were processed with varying amounts of frequency compression, wide dynamic range compression and noise suppression (spectral subtraction). There were interactions between signal processing and voice characteristics as measured using the Hearing Aid Speech Quality Index (HASQI) and listener ratings of perceived sound quality of the processed speech. In this paper, we examine the possible acoustic sources that may explain these interactions, including inter-talker differences in formant space and pitch variation. [Research supported by NIH, GN Resound, NSERC, and the Canada Research Chairs program.]

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Natural differences between talkers in speech production characteristics may affect listeners’ ability to understand talkers (e.g., Bond and Moore, 1994; Ferguson, 2004). Even when intelligibility is unaffected, there may be differences in talkers’ quality of speech as perceived by listeners. These differences may be due to factors inherent to the talker or to changes in the speech signal introduced during processing by technology such as hearing aids. Little research has been done to investigate whether the speech characteristics of talkers may interact with speech processing algorithms, and whether such interactions affect the perceived quality of speech. If such interactions are present, then the implication is that the perceived quality of speech may be improved by tailoring speech processing by hearing aids to the characteristics of talkers.

Voice and speech properties may change with factors such as gender and age. It is well-known that males and females differ on fundamental frequency (F₀) and formant frequencies (Peterson and Barney, 1952). It has also been established that there can be differences between younger and older adults in voice and speech properties. For example, F₀ lowers with age in females, and there may be an increase in the variation of F₃ and amplitude (Mueller, 1997). Speech properties may also change with age, e.g., lowering of F₁ in females (Linville and Fisher, 1985). In addition to F₀ and formant measures, measures such as jitter, shimmer and harmonics-to-noise ratio (HNR) may provide additional information when characterizing the voices of different talkers (Baken and Orlikoff, 2005), and these variations in speech characteristics may interact with speech processing technology to affect perceived quality.

Hearing aid technology has evolved rapidly with advances in digital signal processing and there are many types of processing that alter the speech signal in different ways. Traditionally, the main goal of hearing aids was to amplify speech components to make them audible to listeners with hearing loss while not exceeding the listener’s discomfort level. Today, this goal may be accomplished using syllabic compression. Another goal of hearing aids is to amplify the speech signal while minimizing background noise, which may be achieved using a noise reduction method such as spectral subtraction. Most recently, frequency lowering techniques have been used to restore the availability of high-frequency information to listeners who have high-frequency hearing loss that is not amenable to amplification (e.g., Glista et al., 2009).

There are some indications that different speech processing methods may interact with the acoustical properties of specific voices. In one study, one male and one female talker were ranked differently in listeners’ preferences for a set of five hearing aids (Witter and Goldstein, 1971). In another study with two male talkers and one female talker, the order of listeners’ preferences for hearing aids was different when listening to the two male talkers than when listening to the female talker, and it was speculated that these results were due to differences between talkers in speaking rate, speech spectrum bandwidth or inter-harmonic spacing (Cox and McDaniel, 1984). Formants may also be affected by hearing aid speech processing, e.g., F₁ may be distorted by a low-frequency roll-off (Stelmachowicz et al., 1995), while the height of formant peaks may become flatter relative to the troughs (Bor et al., 2008).

The purpose of this study was to examine whether selected voice and speech properties of talkers would interact with different methods of hearing aid speech processing to affect perceptual quality ratings. Speech samples from five talkers were processed using various hearing aid simulations, and listeners rated the quality of processed samples relative to unprocessed samples.

METHOD

Stimuli

Talker Selection

Talkers were selected from a large database of voices of healthy younger and older adults (Banh et al., 2009). In a previous study, a subset of 60 talkers from this database produced the sentence “People look, but no one ever finds it”, and speech samples were rated by younger and older listeners on pleasantness and the talker’s suitability to be an audiobook reader (Goy et al., 2011). The Hearing-Aid Speech Quality Index (HASQI; Kates and Arehart, 2010) was also calculated for each of these 60 voices in the different processing conditions shown in Table 1. Five talkers were selected so that the set of selected voices encompassed the range of HASQI values that were obtained for the group. These five talkers consisted of one younger and one older adult male and one younger and two older adult females (see Table 2). They were native English speakers from the Southern Ontario region of Canada who reported that they were in good health and had not been diagnosed with any speech or voice disorder. All had pure-tone...
audiometric thresholds of ≤25 dB HL at octave frequencies from 250 to 8000 Hz in the better ear.

Hearing Aid Simulation

The original recordings of the five talkers were re-sampled from 48 kHz to 22 kHz. Sound files were processed in MATLAB using a set of hearing aid simulations (Table 1) and RMS-normalized.

<table>
<thead>
<tr>
<th>Conditions in quiet</th>
<th>Conditions in babble</th>
<th>SNR (dB)</th>
<th>Hearing aid simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>Unprocessed</td>
<td>+ 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unprocessed</td>
<td>− 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spectral subtraction</td>
<td>+ 17.5</td>
<td></td>
</tr>
<tr>
<td>Syllabic compression</td>
<td>Mild + spectral subtraction</td>
<td>+ 17.5</td>
<td>CR = 2; AT = 5 ms; RT = 50 ms; Ch = 18</td>
</tr>
<tr>
<td>Strong</td>
<td>Strong</td>
<td>+ 10</td>
<td>CR = 10; AT = 5 ms; RT = 50 ms; Ch = 18</td>
</tr>
<tr>
<td>Frequency compression</td>
<td>Mild</td>
<td>+ 15</td>
<td>CR = 1.5; CF = 3000 Hz</td>
</tr>
<tr>
<td></td>
<td>Strong</td>
<td></td>
<td>CR = 3; CF = 2000 Hz</td>
</tr>
</tbody>
</table>

TABLE 2. Demographic and perceived speech qualities of five talkers. In a previous study, talkers’ speech samples were rated on a scale of pleasantness from 1 (very unpleasant) to 7 (very pleasant), with 4 as a neutral midpoint, and on a scale to rate the talker’s suitability to be an audiobook reader on a scale from 1 (definitely not) to 5 (definitely yes), with 3 as a neutral midpoint.

<table>
<thead>
<tr>
<th>Talker</th>
<th>Talker ID</th>
<th>Age (years)</th>
<th>Pleasantness</th>
<th>Audiobook reader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Older female</td>
<td>OAF019</td>
<td>70.8</td>
<td>2.6</td>
<td>2.1</td>
</tr>
<tr>
<td>Older female</td>
<td>OAF093</td>
<td>66.8</td>
<td>4.2</td>
<td>2.6</td>
</tr>
<tr>
<td>Older male</td>
<td>OAM072</td>
<td>74.0</td>
<td>5.4</td>
<td>4.1</td>
</tr>
<tr>
<td>Younger female</td>
<td>YAF124</td>
<td>19.3</td>
<td>4.6</td>
<td>2.9</td>
</tr>
<tr>
<td>Younger male</td>
<td>YAM072</td>
<td>18.3</td>
<td>4.8</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Ratings task

Participants

Listeners were 10 normal-hearing native English speakers aged between 21 and 62 years, recruited from the Boulder/Denver metropolitan area. They had pure-tone audiometric thresholds ≤20 dB HL at octave frequencies from 250 to 8000 Hz.

Test Procedure

Listeners were seated in a double-walled sound-attenuating booth. Sentences were presented monaurally through Sennheiser HD 25-1 earphones at an average level of 65 dB SPL, using a Tucker-Davis Technologies System III. Each listener participated in two test sessions, each consisting of five blocks of trials corresponding to the five talkers. In each block, listeners completed one practice trial and three test trials. On every trial, listeners heard 10 processed speech samples and one hidden unprocessed sample from the same talker, following the Multi-Stimulus test with Hidden Reference and Anchor method (ITU, 2003). Listeners were instructed to rate each of the 11 samples relative to a reference (unprocessed) sample on a scale from 0 (bad quality) to 100 (excellent quality), using a sliding scale on a computer interface. The order of these 11 samples was randomized, and listeners could play each sample at their own pace in any order. Over the two test sessions, each listener produced six quality ratings for each of the five talkers, on each of 11 processing conditions. The six ratings were averaged for each listener in each processing condition before further analyses.
Acoustic analysis

Acoustic measures were taken from speech samples using the Praat speech analysis program (Boersma and Weenink, 2012). Three vowels were selected for acoustic analysis: /i/ from “people”, /æ/ from “look” and /i/ from “ever”. Measures of intensity, fundamental frequency (F0), F1, F2, jitter, shimmer and harmonics-to-noise ratio (HNR) were extracted from the steady-state portion of each vowel using a custom script.

RESULTS

Speech Quality Ratings

Perceptual quality ratings by listeners for different talkers across processing conditions are shown in Figure 1. In the quiet conditions, the perceptual quality of the talkers’ speech decreased as the degree of processing increased from mild to strong and it was higher (better) in the syllabic compression conditions than in the frequency compression conditions. In the noise conditions, the quality of talkers’ speech was generally worse than its quality in quiet conditions, and the differences in perceived quality between noise conditions were smaller than the differences between quiet conditions. Quality ratings were near ceiling for the hidden reference condition, while quality ratings for the most adverse noise condition were near floor.

To confirm these descriptions, we conducted a repeated-measures analysis of variance with listeners’ quality ratings as the outcome measure and talker and processing condition as within-subject factors, excluding the two conditions that were near ceiling or floor. There were significant main effects of talker, F(9, 36) = 4.95, p = .003, and processing condition, F(8, 72) = 25.98, p < .001, and a significant interaction of talker with processing condition, F(32, 288) = 2.65, p < .001. Post-hoc t-tests showed that both younger talkers were better than one of the older female talkers (OAF019, p’s < .05), but that the younger talkers were not significantly different from each other, and the older adults were not significantly different from one another. For conditions in quiet, the quality ratings in the two mild processing conditions did not differ significantly from each other, nor did they in the two strong processing conditions. However, quality ratings in the mild and strong syllabic compressions differed from each other, and quality ratings in the mild and strong frequency compression also differed from each other (p’s < .01). For conditions in noise, strong syllabic compression differed on quality ratings from the four other conditions (p’s < .05), which did not differ significantly from each other, except that quality ratings for unprocessed speech at 20 dB SNR were better than ratings for speech processed with mild syllabic compression and spectral subtraction (p < .05).

There were significant interactions between talker and condition, which are summarized in Table 3. For the mild syllabic compression condition in quiet, the only difference between talkers was that the quality of the young male’s speech was judged to be significantly better than that of one of the older females. For the strong syllabic compression condition in quiet, the quality of the young female’s speech was judged to be better than that of all four other talkers. This pattern of results suggests that some property of the younger voices makes them more resistant to the negative effects of syllabic compression when listeners hear them in quiet. For the strong frequency compression condition, the quality of the young male’s speech was judged to be better than that of the young female, while the quality of the older male’s speech was judged to be better than that of one of the older females (OAF093). For the mild frequency compression condition in noise, the quality of the young male’s speech was again better than the young female’s and it was also better than that of one of the older females’ (OAF019), while the quality of the older male’s speech was judged to be better than that of all three females. These results suggest that some property of male voices may underlie their resistance to the negative effects of frequency compression. The final noteworthy pattern in the interactions between talker and conditions was that the speech quality of older female OAF019 was judged to be worse than that of the older male in the mild syllabic with spectral subtraction condition, and it was also worse than that of the younger adults and the other older females in the strong syllabic compression condition. In fact, the quality of the speech of OAF019 was judged to be poorer than that of other talkers in five out of eight conditions in which there were talker differences, suggesting that OAF019 has some particular speech and/or voice properties that interact with a number of hearing aid processing conditions in a deleterious fashion.
The results of the perceptual rating task suggest that younger voices are resistant to distortion by syllabic compression, male voices are more resistant to distortion by frequency compression, and that a voice which already had a poorer perceptual quality before processing (i.e., OAF019) may be more vulnerable to various processing methods.

The mean F0 of vowels remained constant across processing conditions for all talkers (Figure 2). Vowel intensity was decreased relative to unprocessed speech when mild or strong syllabic compression was applied, and this decrease was greater with stronger compression and when vowels were at relatively high amplitude (Figure 3); however, these effects were not associated with any particular talker.

There were three measures related to variability or noise in the voice: jitter, shimmer and HNR. Jitter and shimmer remained stable across processing conditions for all talkers except OAF093, and HNR generally decreased in the syllabic compression conditions (Figure 4).

Strong frequency compression lowered the frequency of F2 for female talkers, but not for male talkers (Figure 5), while strong syllabic compression seemed to raise F1 for all talkers. As shown in Figure 5, male talkers had lower F2’s than female talkers.
When strong syllabic compression was applied, the vowel triangles of the two older female talkers were flattened (with OAF019 being more affected), and the vowel triangle of the older male talker was also distorted. In contrast, the vowel triangles for the two younger talkers remained relatively unchanged when strong syllabic compression was applied.

**FIGURE 2.** Mean $F_0$ of three vowels, /i/, /u/ and /ɛ/, spoken by five talkers in processing conditions in quiet: unprocessed (unproc), mild syllabic compression (MSC), mild frequency compression (MFC), strong syllabic compression (SSC) and strong frequency compression (SFC).

**FIGURE 3.** Mean intensity of three vowels, /i/, /u/ and /ɛ/, spoken by five talkers in processing conditions in quiet: unprocessed (unproc), mild syllabic compression (MSC), mild frequency compression (MFC), strong syllabic compression (SSC) and strong frequency compression (SFC).

**FIGURE 4.** Harmonics-to-noise ratio of three vowels, /i/, /u/ and /ɛ/, spoken by five talkers in processing conditions in quiet: unprocessed (unproc), mild syllabic compression (MSC), mild frequency compression (MFC), strong syllabic compression (SSC) and strong frequency compression (SFC).
syllabic compression. Those talkers who had unusual formants. Younger voices seemed to be less affected by syllabic compression, which may be due to their relatively low formants. Male voices seemed to be less affected by frequency compression, which may be explained in part by formant measures. Male voices remained relatively intact compared to the vowel triangles of older talkers. Those talkers who had unusual formants may be perceived as having a voice that is lower or more nasal.

Female voices were better than one of the older female talkers (OAF019), but the younger talkers were not different from each other and the older talkers were not different from one another. Talker OAF019 was frequently the poorest talker in this study, and she was also the poorest talker in a previous study that did not use a reference voice in the rating task. Although the five talkers included in this study were considered to be normal talkers, they differed in the perceptual quality of their speech after processing by hearing aids. It is likely that such inter-talker differences would be larger in the general population, which includes people with speech or voice pathologies. The relationship between perceptual quality ratings and acoustic properties has been explored in studies with the aim of differentiating normal talkers from those with pathologies, using measures such as harmonics-to-noise ratio (Eskenazi et al., 1990). However, relatively little is known about how perceptual quality is related to the acoustical properties of speech in normal talkers, although there has been some recent interest in how F0 cues are related to various characteristics in normal talkers such as attractiveness (Pisanski and Rendall, 2011) and perceived competence (Klofstad et al., 2012).

In quiet, strong processing was more deleterious than mild processing, but there were no significant differences between syllabic and frequency compression. In general, quality ratings in noise were poorer than quality ratings in quiet, and strong syllabic compression was worse than other conditions tested in noise. Some acoustic measures were generally stable across processing conditions (F0, jitter and shimmer), whereas others were affected by syllabic compression (intensity and HNR). Formants were affected by both syllabic and frequency compression but in different ways; frequency compression had a greater effect on F2 while syllabic compression had a greater effect on F1, which agrees with previous findings that formants may be altered after processing by hearing aids (Stelmachowicz et al., 1995; Bor et al., 2008).

The listeners’ ratings of the talkers’ speech quality depended on both the talker and processing methods; the effects of the interactions between talker and processing method may be explained in part by formant measures. Male voices seemed to be less affected by frequency compression, which may be due to their relatively low formants. Younger voices seemed to be less affected by syllabic compression than older voices, and their vowel triangles remained relatively intact compared to the vowel triangles of older talkers. Those talkers who had unusual vowel triangles in their unprocessed speech (OAM102 and OAM093) showed highly distorted vowel triangles after syllabic compression.

As demonstrated by talker OAF019, talkers with poorer voices to before processing may be perceived as having

**DISCUSSION**

Figure 5. F1 and F2 values of three vowels, /i/, /u/ and /ε/, spoken by five talkers in three conditions in quiet. Also shown are the mean F1 and F2 values for men and women, and the extreme limits of F1 and F2 values for each vowel based on Figure 8 in Peterson and Barney (1952).
a greater decrease in quality relative to other talkers after their speech is processed by hearing aids, even though the extent of distortion of their voice may appear to be similar to that of other talkers when using an objective measure such as the HASQI. When testing the perceptual quality of speech after processing by hearing aids, it may be important to consider the quality of the original signal as well as the changes resulting from hearing aid processing. Finally, ecological validity in the evaluation of the effects of hearing aid processing on speech quality may be increased by the use of a larger variety of voices in addition to those from standardized speech recordings. It may also be useful to extend the corpus of samples spoken by specific talkers to better represent intra-talker differences in speech production that may affect the perceived quality of speech (Muralimanohar et al., 2013). Given the potential for hearing aid processing to adapt to talker differences, these preliminary results suggest that it would be important to conduct further research to determine whether or not such processing would yield sufficient benefit to the listeners to warrant commercial implementation.

ACKNOWLEDGMENTS

This study was funded by the National Institutes of Health, GN ReSound, the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs program. The authors are grateful to Ramesh Muralimanohar and Caleb Kronen for their assistance in stimuli preparation and data collection.

REFERENCES


