Effects of frequency shifts on perceived naturalness and gender information in speech

Peter F. Assmann[1], Sophia Dembling[2] and Terrance M. Nearey[3]

[1, 2] School of Brain and Behavioral Sciences
University of Texas at Dallas, Richardson TX 75083
assmann@utdallas.edu

[3] Department of Linguistics
University of Alberta, Edmonton AB T6G 2E7
t.nearey@ualberta.ca

ABSTRACT

In natural speech, there is a moderate correlation between the fundamental frequency and formant frequencies across talkers. The present study used a high-quality vocoder to manipulate these properties and determine their contribution to perceived naturalness and voice gender. The stimuli were re-synthesized sentences spoken by two adult males and two adult females. Scale factors were chosen for each sentence and for each talker to produce frequency-shifted versions with a specified mean fundamental frequency (F_0) ranging from 60 Hz to 450 Hz in 10 steps, paired with 10 steps in geometric mean formant frequencies ranging from 850 Hz to 2500 Hz. Listeners judged frequency-shifted sentences as more natural when F_0 and formant frequencies followed the co-variation of F_0 and formant frequencies in natural voices. Sentences with low F_0s and low formant frequencies were perceived as masculine, while sentences with high F_0 and high formant frequencies were assigned high ratings of femininity. Sentences with "mismatched" F_0 and formant frequencies were assigned ratings near the midpoint of the range, indicating gender ambiguity. Frequency-shifted sentences derived from male talkers received consistently higher ratings of masculinity than those derived from females and *vice versa*, even when assigned scale factors appropriate for the opposite gender, indicating that factors other than F_0 and mean formant frequencies contribute to perceived gender.
Index terms: frequency-shifted speech, voice gender, perceived naturalness

1. INTRODUCTION

In natural speech, there is a moderate correlation between fundamental frequency (F_0) and formant pattern associated with anatomical differences in laryngeal and vocal tract anatomy across gender and age classes. Figure 1 illustrates the co-variation between F_0 and average formant frequencies (designated here as FF, the geometric mean of F1, F2, and F3) for a sample of 3000 vowels in /hVd/ words (Assmann and Katz, 2000).

Experiments with frequency-shifted vowels have shown that identification accuracy falls dramatically if the formant frequencies are reduced by a scale factor smaller than 0.6 or larger than 1.5 (e.g., Fu and Shannon, 2001). Vowel identification also drops significantly when F_0 is increased or decreased by more than one octave (Assmann et al., 2002). However, Assmann et al. found that an increase in formant frequencies *combined* with an increase in F_0 leads to an improvement in some conditions for vowels spoken by adult males. This synergistic interaction between F_0 and formant pattern was predicted by a model of vowel categorization trained on acoustic measurements of natural vowels spoken by men, women and children. This suggests that listeners may know about the natural co-variation between F_0 and formant pattern and take this into account when identifying vowels and other voiced sounds.

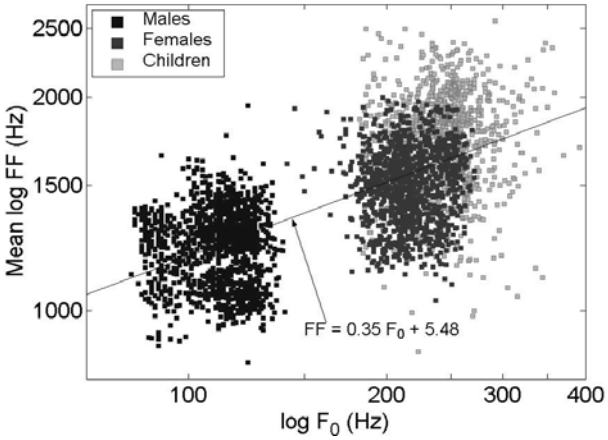

Figure 1: Geometric mean of the formant frequencies (F1, F2, F3) vs. F_0 per vowel token for a sample of vowels spoken by men, women and children.

If listeners have an implicit knowledge of this co-variation, their judgments of the perceived naturalness of frequency-shifted speech should be highest when the resulting speech lies near the regression line in the scatterplot. Alternatively, their responses may reflect a more detailed knowledge of the distribution pattern of F_0 x FF in natural speech.

To test this idea, the present study used a high-quality vocoder called STRAIGHT (Kawahara, 1999) to apply

frequency shifts to a set of recorded sentences. STRAIGHT is a speech analysis-resynthesis system that separates the contribution of source from filter, providing a means of independently shifting F_0 and/or formant pattern up or down along the frequency scale. Formant frequency shifts are implemented in STRAIGHT by shifting the entire spectrum envelope by a multiplicative factor, and thus all formant frequencies are raised or lowered by the same proportion. We used the geometric mean of the lowest three formants (F1-F3) to represent the baseline location of the formant pattern along the frequency scale, a measure related to vocal tract length. We then applied spectrum envelope scale factors to produce a range of mean formant frequencies between 850 and 2500 Hz, spanning the upper and lower extremes in natural speech. In the same way, we generated a set of F_0s between 60 and 450 Hz to reflect the F_0 range in natural speech. The stimuli were presented to two separate sets of listeners. One set made judgments of the perceived gender of the frequency-shifted sentences, while the other group assigned naturalness ratings.

2. METHOD

2.1 Stimuli

Two sentences ("The fly made its way along the wall" and "Two plus seven is less than 10") were spoken by 2 adult males and 2 adult females. F_0 and formant measurements were obtained for all voiced frames in each utterance. Then for each sentence and each speaker, *per-utterance average* F_0 and FF were calculated. Based on these averages, scale factors were chosen so that the same sentence spoken by each of the four talkers (2 males and 2 females) could be assigned *specified* average F_0 and FF values when synthesized using STRAIGHT. Ten F_0 shift factors were chosen for each talker and sentence to produce the following mean F_0s: 60, 75, 94, 117, 147, 184, 230, 288, 360, or 450 Hz. Ten FF shift factors were chosen to produce formant patterns with geometric means of F1, F2, and F3 of 850, 958, 1080, 1218, 1373, 1548, 1745, 1967, 2218, or 2500 Hz.

The manipulated ranges F_0 and FF of the sentences were larger than those we observed in natural data sets when the latter were averaged over all vowels of each speaker. These were vowels spoken by men, women, and children aged three years and older (Hillenbrand et al., 1995; Assmann and Katz 2000). The manipulated F_0 ranged from approximately 0.74 times the minimum to 1.17 times the observed maximum, while the manipulated FF ranged from about 0.82 times the minimum to 1.36 times the observed maximum.

2.2 Listeners

Listeners were 30 undergraduate students in Psychology who participated for partial course credit. All were native speakers of American English and had normal hearing based on a hearing screen and self-report. None had participated in previous experiments listening to synthesized speech.

2.2 Procedure

The stimuli were presented diotically over headphones in a double-walled sound booth. All conditions (talkers, sentences, F_0 shifts, and FF shifts) were randomly interspersed. Listeners were informed that they would hear a range of computer-generated voices varying in naturalness, and that some voices might sound like children or cartoon characters. Listeners rated the voice using a graphical slider displayed on the computer screen. For the naturalness judgments, the extreme left position of the slider was labeled 'highly unnatural', while the extreme right position was labeled 'definitely natural'. Intermediate positions were labeled 'slightly unnatural' and 'possibly natural'. For the gender ratings listeners were instructed to indicate whether they heard the voice as masculine or feminine. The label 'clearly masculine' was displayed on the extreme left position; 'clearly feminine' was displayed on the right. Equally spaced between these two extremes were four intermediate settings labeled as 'somewhat masculine', 'slightly masculine', 'slightly feminine', and 'somewhat feminine'.

3. RESULTS

3.1 Naturalness Judgments

Figure 2 shows that naturalness varies systematically as a function of F_0 (upper panel) and formant frequency (lower panel). Averaged across all formant frequency shifts, sentences with F_0s in the adult female range (233 Hz) were reported as more natural than vowels with lower or higher F_0s, $F(9,81)=28.35$; $p<0.01$. A similar pattern is seen for formant frequency shifts averaged across all F_0 conditions, with the highest mean naturalness rating assigned to step 7 along the FF scale (corresponding to a geometric mean of F1-F3 of 1950 Hz, also near the middle of the adult female range), $F(9,81)= 116.03$; $p<0.01$. Compared to F_0 shifts, the variation in naturalness as a function of formant frequency shifts was smaller and the overall pattern was flatter. However, formant frequency explained a larger proportion of the variance in listeners' judgments of naturalness than F_0 ($r^2=0.65$ for formant frequency shifts, compared to $r^2=0.35$ for F_0 shifts).

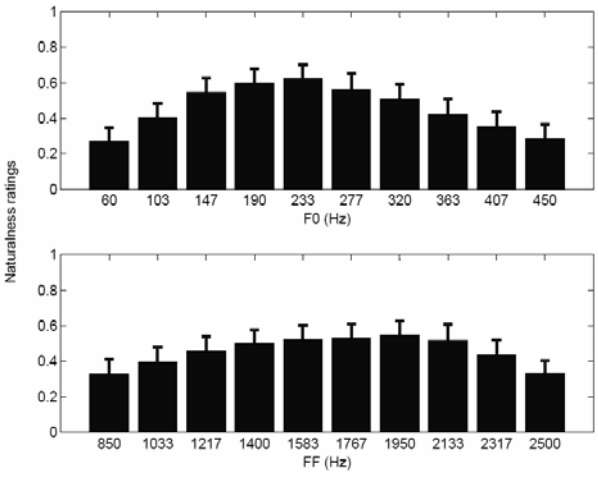

Figure 2: Mean naturalness ratings as a function of shift in F_0 (averaged across FF, upper panel) and FF (averaged across F_0, lower panel).

Naturalness ratings varied systematically as a function of the degree of match between F_0 and FF, giving rise to a significant interaction, $F(81,729)=30.81$; $p<0.01$. The interaction is shown in Figure 3, in which perceived naturalness is shown by the size of the circles, with larger circles reflecting a higher degree of naturalness. Matched combinations (i.e., low F_0s combined with low FF, or high F_0s paired with high FF) were judged as more natural than mismatched combinations. Overall, sentences judged as more natural occupied the region near the regression line, except at the highest and lowest extremes. In general, the highest naturalness ratings were assigned to sentences with F_0 x FF combinations that could be found in natural vowels. A fairly strong correlation was found (across the 100 shift conditions) between naturalness ratings and distance from the regression line (r=0.83).

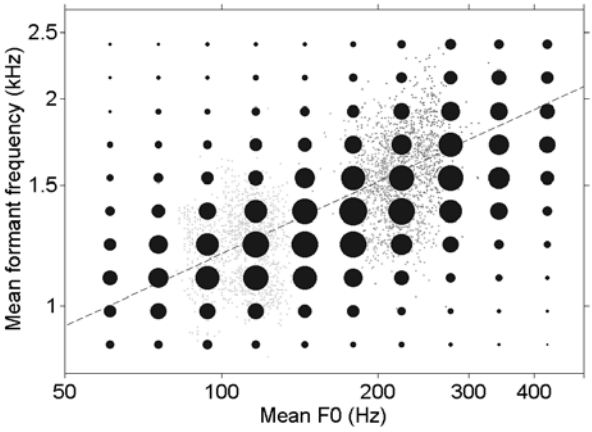

Figure 3: Interaction of F_0 and FF shifts in naturalness judgments. Circle size is proportional to the mean naturalness rating assigned by listeners (N=10). For comparison, the small dots show the measurements of natural vowels presented in Figure 1.

3.2 Voice Gender

Mean gender ratings are shown in Figure 4 by the size of the circles, which expresses the deviation in slider position from the "neutral" or midpoint setting. The large gray circles indicate slider settings approaching the maximum position in the “extremely masculine” direction. The large black circles indicate ratings in the “extremely feminine” direction. Combinations of low F_0 with low FF were judged as strongly masculine, and combinations of high F_0 with high FF were perceived as feminine. In general, the ratings of masculinity resulting from downward shifts were higher than the ratings of femininity produced by upward shifts. F_0 and FF combinations that were mismatched (e.g., low F_0 and high FF, or *vice versa*) were assigned rankings near the middle of the range (indicating gender ambiguity).

Further analysis of the data revealed an interaction that was not apparent in the combined data shown in Figure 4. Figure 5 presents the data separately for the subset of sentences that were originally spoken by males (upper panel) or females (lower panel). Compared to Figure 4, in the upper panel the gray circles are on average larger, and occupy a larger area than the black circles, indicating larger area than the black circles, indicating consistently higher

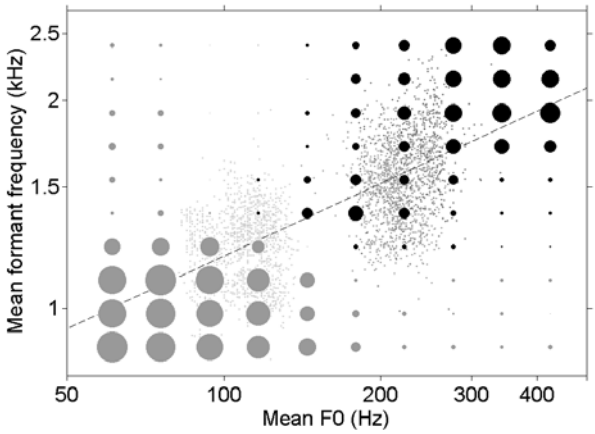

Figure 4: Interaction of F_0 and FF shifts in gender ratings. Gray circles indicate ratings in the direction of masculine voices, with larger circles representing a higher degree of masculinity. Black circles indicate ratings in the direction of feminine voices.

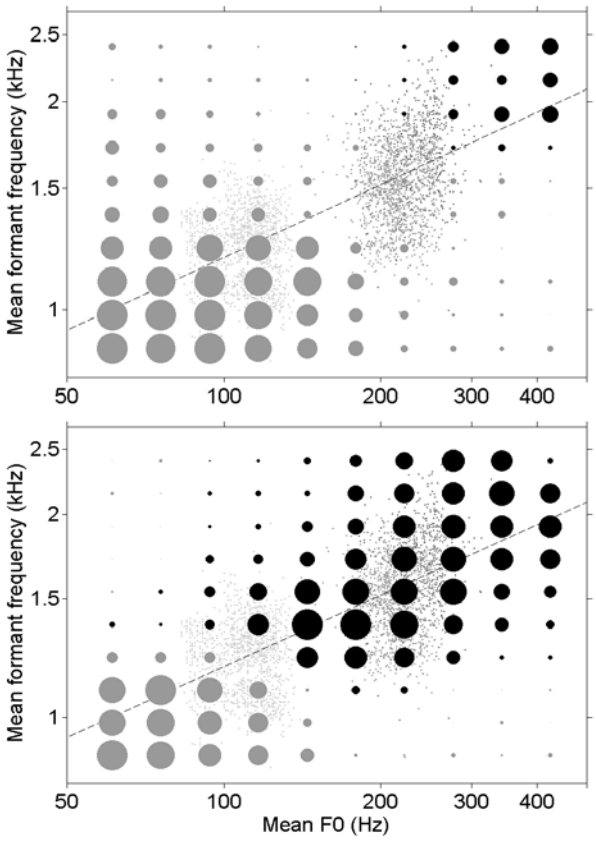

Figure 5: Same as Figure 4, but showing the subset of sentences originally spoken by male talkers (upper panel) or female talkers (lower panel).

ratings of masculinity. Similarly, the lower panel shows a corresponding increase in the ratings of femininity. Thus sentences originally spoken by male talkers received higher ratings of masculinity, and sentences spoken originally spoken by females received higher ratings of femininity, regardless of the assigned F_0 and formant pattern. A plausible interpretation is that factors other than mean F_0 and formant pattern contribute to the perception of voice gender.

4. DISCUSSION

Taken together, the results indicate that listeners take account of the co-variation of F_0 and FF in natural speech when judging the perceived naturalness and voice gender of frequency-shifted sentences. Our findings are consistent with recent work by Hillenbrand (2005) who found that upward shifts in F_0 and formant frequencies of male voices were usually effective in changing the perceived gender from male to female, while downward shifts applied to female voices had the opposite effect. However, he concluded that male-female judgments must be determined in part by additional cues, because approximately 20% of the utterances remained unchanged even when F_0 and formant frequencies were shifted to values appropriate for the opposite sex.

Smith and Patterson (2005) reported speaker size and sex/age judgments (man, woman, boy, girl) for vowels that were scaled using STRAIGHT. They found that glottal pulse rate and estimated vocal tract length (corresponding to F_0 and FF in our terminology) made roughly similar contributions to judgments of speaker sex/age for F_0 and FF in the normal range, while FF information was the determining factor for mismatched voices with low F_0 and high FFs. In comparison, our results suggest roughly equal contributions of F_0 and FF for matched voices and a high degree of gender ambiguity (and reduced naturalness) for mismatched voices. Smith and Patterson found that voices with low F_0 and high FF (upper left quadrant of Figure 4) were predominantly classified as boys. However, our results show that voices with these combinations of F_0 and FF are assigned gender ratings near the middle of the scale, indicating gender ambiguity. Differences in the stimuli (vowels vs. sentences) and in the nature of the task (sex/age classification vs. rating scale) may account for this difference in results.

Smith and Patterson found relatively low overall response probabilities for the category "woman", consistent with our finding that listeners assigned lower overall ratings of femininity to voices in the upper right quadrant of Figure 4. They attribute this to overlap in natural distributions of F_0 and FF from women, boys, and girls. Gender classification is generally less accurate for children's voices than for adults (Perry et al., 2001) and this may lead to greater uncertainty about voice gender with upward scaling of F_0 and FF.

Our findings extend Smith and Patterson's findings in two respects. First, they asked listeners to make categorical judgments of speaker sex/age, whereas listeners in our study provided graded judgments along a masculine-feminine scale. Second, they used only a single male talker producing vowels, whereas we used both male and female talkers producing meaningful sentences. Our results suggest that there are residual indicators of voice gender in these sentences, even when they undergo substantial shifts in F_0 and FF. Further research is needed to determine the basis for these effects.

5. ACKNOWLEDGEMENTS

We thank Tiffani Jantz for assistance in conducting the experiments and Hideki Kawahara for providing the STRAIGHT synthesis software. Portions of this work were reported at the 149th Meetings of the Acoustical Society of America. This work was supported in part by grant #0318451 from the National Science Foundation.

6. REFERENCES

Assmann P.F. and Katz, W.F. "Time-varying spectral change in the vowels of children and adults," *J. Acoust. Soc. Am*. 108, pp. 1856-1866, 2000.

Assmann P.F., Nearey T.M., and Scott J.M., "Modeling the perception of frequency-shifted vowels," *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 425-428, 2002.

Fu Q-J. and Shannon R.V., "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," *J. Acoust. Soc. Am*., 105, pp. 1889-1900, 1999.

Hillenbrand J.M, Getty L.A., Clark, M.J., and Wheeler, K. "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am., 97,* 3099-3111, 1995.

Hillenbrand J.M., "The role of fundamental frequency and formants in the perception of speaker sex," *J. Acoust. Soc. Am*., 118, pp. 1932-1933, 2005.

Kawahara H., "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," *Proceedings of the ICASSP*, pp. 1303-1306, 1997.

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

Smith, D.R. and Patterson, R.D. "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex and age," *J. Acoust. Soc. Am*., 118, pp. 3177-3186, 2005.