THE ROLE OF VOICE SOURCE MEASURES ON AUTOMATIC GENDER CLASSIFICATION

Yen-Liang Shue and Markus Iseli

University of California Los Angeles
Dept. of Electrical Engineering
405 Hilgard Ave., Los Angeles, CA 90095
yshue@ee.ucla.edu, iseli@ee.ucla.edu

ABSTRACT

Differences of physiological properties of the glottis and the vocal tract are partly due to age and/or gender differences. Since these differences are reflected in the speech signal, acoustic measures related to those properties can be helpful for automatic age and gender classification. In this paper, the focus is on the role of acoustic measures related to the voice source in automatic gender classification, implemented using Support Vector Machines (SVMs). Acoustic measures of the vocal tract and the voice source were extracted from 3880 utterances spoken by 205 male and 160 female talkers (aged 8 to 39 years old). Formant frequencies and formant bandwidths were used as vocal tract measures, and open quotient and source spectral tilt correlates were used as voice source measures. Results show that the addition of voice source measures can help improve automatic gender classification results for most age groups.

Index Terms— voice source, gender classification, gender identification

1. INTRODUCTION

Gender-based differences in human speech are due in part to physiological differences such as vocal fold thickness or vocal tract length, and differences in speaking style. Physiological properties of the glottis and the vocal tract change with age and gender. Since these changes are reflected in the speech signal, acoustic measures related to those properties can be helpful for age and gender classification. Assuming the linear source-filter model of speech production [1], the contribution of acoustic measures to such classification can then be attributed to the voice source or the vocal tract. To our knowledge, with the exception of fundamental frequency (F_0), there has been no study that has examined the role of measures related to the voice source on age and/or gender classification.

It is well known that F_0 values for male talkers drop during adolescence due to a lengthening and thickening of the vocal folds. F_0 for adult males is typically around 120 Hz, while F_0 for adult females is around 200 Hz [2]. This effect is mostly due to a lengthening and thickening of the male vocal folds.

It is also well known that, due to vocal tract length differences, adult males exhibit lower formant frequencies than adult females [2]. Interestingly, for preadolescent children, studies also found lower formant frequencies for boys compared to girls of ages 5-6 [3], 7-8 years [4], and ages 5, 7, 9, and 11 years (for Australian English) [5]. These findings imply that, overall, boys have larger vocal tracts than girls. In [6], statistical analysis of children speech confirmed that formant frequencies (F_1, F_2, F_3), and not F_0, differentiate gender for children as young as 4 years of age, while formant frequencies plus F_0 differentiate gender after 12 years of age. These findings lead to the conclusion that for preadolescent children, vocal tract measures play a bigger role for gender classification than the voice source measure F_0. For adult speech, automatic gender classification has been presented in [7], which used linear predictive coding (LPC)-derived measures that represent the vocal tract.

In [8], changes in magnitude and variability of, among other measures, F_0, formant frequencies, and spectral envelope are presented as a function of age for talkers from 5 to 50 years old. For F_0, the study showed a drop between ages 12 and 15 for males and a drop of F_0 variation for all talkers between ages 5 and 15. Formant frequencies (F_1, F_2, F_3) decreased between ages 10 and 15, where formant frequencies of male talkers decreased faster and reached much lower absolute values than those of female talkers. The study showed that children younger than age 10 displayed greater spectral variability than adults.

In [9], we analyzed age, sex, and vowel dependencies, for talkers between the ages of 8 and 39, of the following three voice source measures: F_0; $H_1^* - H_2^*$, the difference of the first two source spectral harmonic magnitudes (related to the open quotient[1] [10]); and $H_1^* - A_3^*$, the difference of the first source spectral harmonic magnitude and the magnitude of the source spectrum at the frequency location of the third formant (related to source spectral tilt [10]). The asterisk indicates a correction for the influence of vocal tract resonances [11]. For male talkers, the results showed a drop of about 5 dB in $H_1^* - H_2^*$ around age 15 and a continuous decrease of $H_1^* - A_3^*$ between ages 8 and 39 by about 10 dB. For female talkers, the value of $H_1^* - H_2^*$ remained relatively unchanged between ages 8 and 39, whereas for $H_1^* - A_3^*$ a slight decrease by about 4 dB was shown. These developmental changes resulted in higher values of F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$ for adult female talkers compared to adult male talkers [12].

In this paper, acoustic measures from both the voice source and the vocal tract are used for automatic gender classification of 8 to 39 year old talkers. The vocal tract measures consist of formant frequencies and formant bandwidths, and the voice source measures used are F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$. Training and testing is done using support vector machines (SVMs). The results are analyzed to see if voice source measures can improve automatic gender classification. Finally, the SVM classification results are compared with human perception classification tests, and also with classification results using conventional Mel-frequency cepstral coefficient (MFCC) features in combination with Gaussian mixture models (GMMs).

[1]The open quotient is defined for voiced speech as the ratio between the glottis open time and the fundamental period.

2. SPEECH DATA

Speech recordings from five age groups, ages 8–9, 10–11, 12–13, 14–15 and 16–39 were taken from the CID database [13]. Each recording was of the form "I say uh, bVt again", where the target vowel 'V' was /ih/, /eh/, /ae/ or /uw/. The vowel /iy/ in 'bead' was also used. These utterances were spoken at the habitual speaking level and most talkers repeated the phrases twice. For the analysis, only the manually segmented target vowels were used. The distribution of talkers (males/females) and number of utterances per age group is listed in Table 1. The total number of male/female talkers is 205/160 and the total number of utterances is 3880.

Table 1. *Distribution of gender and utterances for each age group.*

Age group	males/females	No. of utterances
8-9	48/36	810
10-11	48/33	807
12-13	38/34	708
14-15	22/21	413
16-39	49/36	1142

3. METHODS

The acoustic measures used for gender classification were the first three formant frequencies (F_1, F_2, and F_3), the first two formant bandwidths (B_1 and B_2), and the measures related to the voice source F_0, $H_1^* - H_2^*$, and $H_1^* - A_3^*$. The third formant bandwidth, B_3, was not used due to its large variance. The formant frequencies and bandwidth values were estimated using the "Snack Sound Toolkit" software [14] with these settings: analysis window length of 25 ms, window shift of 1 ms and pre-emphasis factor of 0.96. F_0 was extracted using the STRAIGHT algorithm [15]. The spectral magnitudes H_1, H_2, and A_3 were estimated from the speech spectrum using the values of F_0 and F_3. Corrections, denoted by the asterisks, were made to these measures to remove the effects of the vocal tract[11]. For each of the voice source measures, a first order Legendre polynomial was fitted to the raw values to obtain a measure of the mean and the slope (denoted by $\triangle$) across the duration of the vowel.

Classification was done using an SVM classifier with a Radial Basis Function kernel. In this study, the LIBSVM toolkit [16] was used to train and test on vectors containing different combinations of acoustic measures extracted from the five target vowels. For each classification experiment, 70% of the utterances, selected randomly, were used for training; the remaining utterances were used for testing. Five experiments were performed for each combination of acoustic measures and the average accuracy recorded.

For perception tests, four male subjects between ages 26 and 39 participated. They were each presented with 100 utterances of the target words and had to decide between male or female voice. The target words were manually segmented from the carrier phrase and were played back in random order using headphones. The distribution of male and female utterances per age group are listed in Table 2. The same perception tests were also performed using just the segmented vowel part of the target word.

To compare the SVM results with more traditional methods, the first 12 MFCCs were extracted from the utterances and combined with the mean F_0 for each of the utterances to form a 13-dimension feature vector. Training was done with 2 GMMs each with 6 mixtures.

Table 2. *Distribution of utterances used in perception experiments.*

Age group	No. of utterances male/female
8-9	7/7
10-11	8/8
12-13	8/8
14-15	12/10
16-39	15/17

4. RESULTS AND DISCUSSION

For this section, the set of acoustic measures containing formant information (F_1, F_2, F_3, B_1, and B_2) will be denoted by FB.

4.1. Results using F_0 and formants

As a first step, we analyzed the contribution to gender classification accuracy of only F_0, only FB, and F_0 plus FB (labeled by M0). These measures are the most widely used in gender and age classification. Figure 1 shows the classification accuracy for each age group using those measures. For ages 8 to 11 it can be seen that formant information only (FB) performs slightly better than F_0. This is consistent with [6]. Gender classification accuracy for ages 8 to 13 is always below 65%, but between age groups 12–13 and 14–15, it increases to 85% for F_0 and to 68% for FB; these results can be attributed to the large drop of F_0 for males around ages 12 to 15 (about 105 Hz on average) [9, 8] and to a decrease of formant frequencies for males relative to females [8]. Since M0 overall yielded the best results, it was chosen as the baseline measure set for the comparison of the performance of voice source measures.

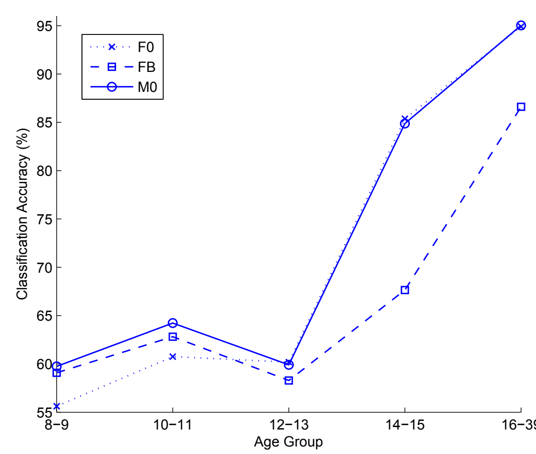

Fig. 1. Gender classification accuracy for each age group using just F_0, just FB, and F_0 plus FB (M0).

Table 3. *Measure sets (M0-M3) used in the gender classification tests. M0, in bold, is used as the baseline measure set.*

Set	Acoustic Measures					
	$F0$	FB	$H_1^* - H_2^*$	$H_1^* - A_3^*$	$\triangle F_0$	$\triangle H_1^* - H_2^*$
M0	✓	✓				
M1	✓	✓	✓			
M2	✓	✓	✓	✓		
M3	✓	✓	✓	✓	✓	✓

4.2. Results adding voice source measures

Figure 2 compares the changes in gender classification accuracies resulting from the addition of the various voice source measure sets (M1–M3) as listed in Table 3. The baseline measure set (M0) is shown as a solid line. Table 4 shows the values corresponding to this figure as well as results from MFCC/GMM classification tests. It can be seen that adding voice source measures plays a significant role only for age groups 10–11 and 12–13, where the absolute accuracy was improved by up to 9% using measure set M3. For age group 8–9, the accuracies are below 60% and the SVM seems unable to model the classes for males and females satisfyingly. Although it was shown in [9] that the source measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$ are dependent on age and gender, the changes in classification accuracy for age groups 14–15 and 16–39 when using M1 or M2 are not significant. This could be attributed to the already large classification accuracy of the baseline (M0). Interestingly, while the classification accuracies for the voice source measure sets are similar to the MFCC/GMM results for age groups 8–9, 12–13 and 16–39, the voice source measure set performance for M2 is about 9% and 5% higher for age groups 10–11 and 14–15, respectively.

A closer look at the classification accuracy results for age group 12–13 is shown in Table 5, which shows the percentage correct classification of males and females. Compared to M0, the addition of the voice source measures assists in increasing the classification accuracy by about 7% for males and 9% for females when using M3. However, since the M2 measures are easier to calculate than those of M3, and M2 showed a classification accuracy improvement for all ages between 10 and 39, it is recommended to use M2 for gender classification. M2 will be used throughout the remainder of this paper.

Table 4. *Gender classification accuracy for the different measurement sets (M0-M3) and age groups. MFCC feature classification results are shown for comparison.*

Age group	Baseline set M0	Voice source measure sets M1	M2	M3	MFCC features
8-9	59.75%	58.76%	58.18%	59.83%	59.01%
10-11	64.23%	64.07%	67.30%	65.39%	58.34%
12-13	59.91%	63.51%	65.50%	68.63%	68.91%
14-15	84.88%	86.50%	86.18%	82.93%	81.63%
16-39	95.03%	95.26%	95.15%	94.85%	95.79%

4.3. Comparison with perception results

Table 6 compares automatic classification results (denoted by AUT) with human perception results from this study (denoted by PER1) and from perception experiments in [6] (denoted by PER2). Note in [6], the target words were in a different context (hVd instead of bVt). These perception experiments were done using the target words. All values are gender recognition accuracies in percent. Dashes in the table represent unavailable data. AUT results were using measure set M2. The SVM classifier performs comparably with the human subjects for the talkers aged 14 and above. For talkers aged below 14, the results are somewhat mixed and the accuracies reduce with decreasing age; however this trend also exists with the human classifiers. In effect, in the "total" section of the table, the AUT results agree well with the perception results.

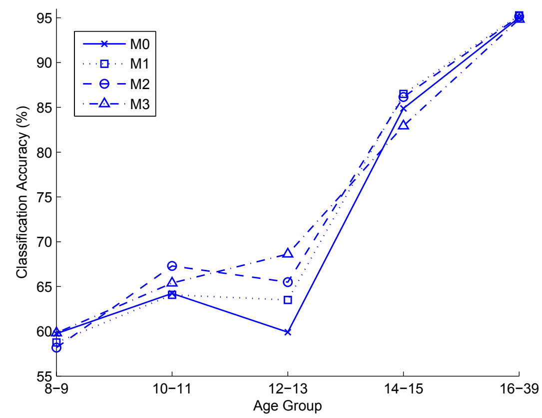

Fig. 2. Gender classification accuracy for each age group using the measures sets M1, M2 and M3. M0 represents the baseline performance results. The corresponding values are listed in Table 4.

Table 5. *Gender classification accuracy for age group 12-13, distinguishing between males and females.*

Set	M	F	Total
M0	59.28%	60.60%	59.91%
M1	63.24%	63.80%	63.51%
M2	63.06%	68.20%	65.50%
M3	66.67%	70.00%	68.63%

Since the SVM was only given the target vowels, and the listeners were able to listen to the whole target word, it seemed only fair to see how listeners would perform when given only short vowel segments. Interestingly, for talkers of age 15 and above, the results were similar to gender classification using target word (about 90% recognition accuracy) and our experimental subjects were mostly using F_0 to do the classification. For talkers of age 14 and below however, our experimental subjects all agreed that their decisions were on target vowels were mostly based on chance; the removal of the contextual information reduced the ability to distinguish between genders. As stated in [6]: "...prosodic features that are overlayed (suprasegmentals) upon sound segments in words, phrases, or sentences and include intonation, stress, duration, and juncture maybe important in gender identification."

5. SUMMARY AND CONCLUSIONS

In this paper, we examined the role of voice source measures in automatic gender recognition and compared the results to perceptual

Table 6. *SVM gender classification accuracy, in percent, using measure set M2 compared with perception results from this paper (PER1) and from Perry et al. [6](PER2). Dashes indicate unavailable values. The perception experiments used the target words.*

Age	8	9	10	11	12	13	14	15	16
	Males								
AUT	-		-		67		83		94
PER1	39		72		91		100		100
PER2	74	-	-	-	82	-	-	-	99.7
	Females								
AUT	-		-		68		90		97
PER1	68		75		31		70		97
PER2	56	-	-	-	56	-	-	-	95
	Total								
AUT	58		67		66		87		95
PER1	54		73		61		86		98
PER2	65	-	-	-	69	-	-	-	97

experiments performed on the same database. Vocal tract and voice source measures were extracted from a large database of 3880 utterances spoken by 205 males and 160 females. Formant frequencies and formant bandwidths were used as vocal tract measures, and F_0, $H_1^* - H_2^*$ (related to open quotient), and $H_1^* - A_3^*$ (related to spectral tilt) were used as voice source measures. The slopes (derivatives) were also calculated for the voice source measures. Automatic gender classification using SVMs was performed on five age groups with different sets of acoustic measures.

Using a baseline measure set consisting of F_0, the first three formants (F_1, F_2, F_3) and the first two bandwidths (B_1, B_2), it was found that adding the two voice source measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$ yielded the most consistent classification accuracy improvement over the baseline. For age group 8–9, the results were all below 60%, slightly higher than chance, however for ages greater than 9, using these two measures increased the classification accuracy, although the improvements decreased for older talkers as the role of F_0 became more dominant. The measure sets which included the slopes $\triangle F_0$ and $\triangle H_1^* - H_2^*$ did not produce consistent results and in some age groups actually reduced the classification accuracy.

Perception experiments using the target words showed similar results compared to the results of the SVM classifier, which used only the target vowel. Perception experiments using only the target vowel showed that for children aged 14 and below, classification accuracy was close to chance, suggesting that outside the vowel segment there exist suprasegmental cues, which could aid in automatic gender classification. Future work will focus on finding reliable methods to extract these suprasegmental cues.

6. REFERENCES

[1] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, Paris, 1960.

[2] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, March 1952.

[3] B. Weinberg and S. Bennett, "Speaker sex recognition of 5- and 6-year-old children's voices," *The Journal of the Acoustical Society of America*, vol. 50, pp. 1210–1213, 1971.

[4] S. Bennett, "Vowel formant frequency characteristics of preadolescent males and females," *The Journal of the Acoustical Society of America*, vol. 69, pp. 231–238, 1981.

[5] P. Busby and G. Plant, "Formant frequency values of vowels produced by preadolescent boys and girls," *The Journal of the Acoustical Society of America*, vol. 97, pp. 2603–2606, 1995.

[6] T. L. Perry, R. N. Ohde, and D. H. Ashmead, "The acoustic bases for gender identification from childrens voices," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2988–2998, June 2001.

[7] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[8] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, March 1999.

[9] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustical measures related to the voice source," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2283–2295, April 2007.

[10] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *J. Speech Hear. Res.*, vol. 38, pp. 1212–1223, 1995.

[11] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proceedings of ICASSP*, Montreal, Canada, May 2004, vol. 1, pp. 669–672.

[12] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *The Journal of the Acoustical Society of America*, vol. 106, pp. 1064–1077, 1999.

[13] J.D. Miller, S. Lee, R.M. Uchanski, A.F. Heidbreder, B.B. Richman, and J. Tadlock, "Creation of two children's speech databases," in *Proceedings of ICASSP*, May 1996, vol. 2, pp. 849–852.

[14] Kåre Sjölander, "Snack sound toolkit," KTH Stockholm, Sweden, 2004, http://www.speech.kth.se/snack/ (last viewed Oct. 2007).

[15] H. Kawahara, A. de Cheveign, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised TEMPO in the STRAIGHT-suite," in *Proceedings ICSLP'98*, Sydney, Australia, December 1998.

[16] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm (as of Oct. 2007).