

# ON USING VOICE SOURCE MEASURES IN AUTOMATIC GENDER CLASSIFICATION OF CHILDREN'S SPEECH

Gang Chen, Xue Feng, Yen-Liang Shue, and Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles

gangchen@ee.ucla.edu, xfeng@ucla.edu, yshue@ee.ucla.edu, alwan@ee.ucla.edu

## Abstract

Acoustic characteristics of speech signals differ with gender due to physiological differences of the glottis and the vocal tract. Previous research [1] showed that adding the voice-source related measures  $H_1^* - H_2^*$  and  $H_1^* - A_3^*$  improved gender classification accuracy compared to using only the fundamental frequency ( $F_0$ ) and formant frequencies.  $H_i^*$  refers to the  $i$ -th source spectral harmonic magnitude, and  $A_i^*$  refers to the magnitude of the source spectrum at the  $i$ -th formant. In this paper, three other voice source related measures:  $CPP$ ,  $HNR$  and  $H_2^* - H_4^*$  are used in gender classification of children's voices.  $CPP$  refers to the Cepstral Peak Prominence [2],  $HNR$  refers to the harmonic-to-noise ratio [3], and  $H_2^* - H_4^*$  refers to the difference between the 2nd and the 4th source spectral harmonic magnitudes. Results show that using these three features improves gender classification accuracy compared with [1].

**Index Terms:** gender classification, gender identification, voice source

## 1. Introduction

Previous studies on automatic gender classification from speech signals of adult speakers achieved high accuracy by using only features related to the fundamental frequency ( $F_0$ ) and the first four formant frequencies [4]. This is mainly due to the well-known physiological differences between adult male and female speakers. However, automatic gender classification from speech signals for children and adolescents remains a challenge because  $F_0$  and formant frequencies are not easily distinguishable between boys and girls.

Existing studies of children's voices have mainly focused on the formant properties. In [5], the voices of children between the ages of 5 and 11 years old were studied. By using target words, which represented non-diphthong vowels in Australian English, the study was able to show that the value of the first three formant frequencies for girls were higher than those for boys, while boys have higher formant amplitudes than girls. In [6],  $F_0$ , formant frequencies and measures relating to the spectral envelope were studied as a function of age for speakers between ages 5 and 50. That study showed that the  $F_0$  value dropped between ages 12 and 15 for males, and formant frequencies decreased between ages 10 and 15. With increasing age, male speakers also showed a faster decrease in formant frequencies than female speakers, and the formant frequencies after the decrease were lower for male than for female speakers. In [7], speech from children of ages 4, 8, 12 and 16 were studied; each age group consisted of 10 boys and 10 girls. An analysis of seven non-diphthong vowels of American English showed that the formant frequencies differentiated gender before 12 years of age, while formant frequencies along

with  $F_0$  differentiated gender after 12 years of age. These studies suggest the usefulness of pitch as a distinguishing feature diminishes as the differences between  $F_0$  for the two genders decrease.

Although vocal-tract related features, including formant frequencies and their amplitudes, have been studied to differentiate gender, the role of voice-source related measures in gender classification have not been systematically investigated. The effects of age, sex and vowel dependencies on some measures related to the voice source were analyzed in [8]. The measures were analyzed from the speech data of speakers between the ages of 8 and 39, and included:  $F_0$ ,  $H_1^* - H_2^*$ , the difference between the first two source spectral harmonic magnitudes (related to the open quotient [9]), and  $H_1^* - A_3^*$ , the difference between the first source spectral harmonic magnitude and the magnitude of the source spectrum at the frequency location of the third formant (related to spectral tilt [9]). The asterisk indicates a correction for the influence of vocal tract resonances using the formula given in [10]. Results showed that  $H_1^* - A_3^*$  continuously decreases between ages 8 and 39 by about 10 dB for males and decreases slightly by about 4 dB for females. It also suggested that  $H_1^* - H_2^*$  drops about 5 dB at around age 15 for males but remains relatively unchanged for females. These differences motivated the study in [1] where acoustic measures from both the voice source and the vocal tract were used for automatic gender classification for speakers aged 8 to 39. It was found that the addition of two measures,  $H_1^* - H_2^*$  and  $H_1^* - A_3^*$ , yielded the most consistent classification accuracy improvement when compared to the baseline ( $F_0$  and formant frequencies). The results suggested that voice source measures could contain discriminative information for gender classification.

In [11], two sentences by ten female and six male talkers were analyzed and results showed that, on average, females are more breathy than males among English speakers.  $CPP$  is defined in [2] as "a measure of the amplitude of the cepstral peak corresponding to the fundamental period, normalized for overall signal amplitude". A signal with well defined periodic structure is expected to show a very prominent cepstral peak. Hence,  $CPP$  has been used to differentiate between breathy signals and nonbreathy signals. In [2], the effectiveness of several acoustic measures in predicting breathiness was evaluated. Perceptual tests were conducted to obtain breathiness ratings from a sustained vowel and a 12-word sentence spoken by 20 speakers with voice pathologies and 5 speakers with no voice pathologies. Results showed that  $CPP$  is highly correlated with breathiness ratings.

Harmonic-to-noise ratio ( $HNR$ ) is a measure of harmonic energy normalized by the spectral noise level [3]. In [3], the sensitivity of  $HNR$  to jitter was tested with synthetic vowel-like

signals. Results indicated the strong negative relation between  $HNR$  and jitter. Since  $HNR$  indicates the noise level and [11] showed that listeners were more likely to rate a signal as being breathy if random noise was added to the signal along with an increase in  $H_1$ ,  $HNR$  could be an indicator of breathiness.

In this paper, additional measures relating to the voice source, such as  $CPP$ ,  $HNR$  and  $H_2^* - H_4^*$ , are extracted and used in conjunction with  $F_0$  and formant values for automatic gender classification. Results are compared with those in [1] as well as with Mel-frequency cepstral coefficient (MFCC) features.

## 2. Data

Speech data are from the CID database [12] produced by five age groups: ages 8–9, 10–11, 12–13, 14–15 and 16–17. Each recording was of the form: “I say uh, bVt again”, where the vowel ‘V’ was /ih/, /eh/, /ae/ or /uw/. The vowel /iy/ in ‘bead’ was also used. Each speaker had, on average, 20 utterances of this form with different vowels. Table 1 shows the distribution of the utterances in terms of gender and age groups. The total number of male/female speakers is 174/140, and the total number of utterances is 3418. The steady state part of each vowel was extracted manually for analysis.

Table 1: Distribution of utterances in terms of gender and age.

Age group	No. of 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
16–17	18/16	680

## 3. Methods

The vocal tract parameters used in this study were the first three formant frequencies ( $F_1$ ,  $F_2$  and  $F_3$ ) and the formant bandwidths ( $B_1$ ,  $B_2$ ). Also used were measures related to the voice source:  $F_0$ ,  $CPP$  (related to breathiness [2]),  $HNR$  (the harmonic-to-noise ratio [3]),  $H_1^* - H_2^*$ ,  $H_1^* - A_3^*$ , and  $H_2^* - H_4^*$  (the difference between the second and fourth source spectral harmonic magnitudes; related to mid-frequency tilt [13]). Additional measures used are amplitudes of first three formant frequencies ( $A_1$ ,  $A_2$  and  $A_3$ ).

$HNR$  was calculated in the frequency band of 0–3500 Hz. The formant frequencies were estimated using the “Snack Sound Toolkit” software [14] using the following settings: window length 25 ms, window shift 1 ms and pre-emphasis factor of 0.96.  $F_0$  values were estimated using the STRAIGHT algorithm [15]. The harmonic magnitudes,  $H_1^*$ ,  $H_2^*$  and  $H_4^*$ , were calculated from the speech spectrum using the  $F_0$  values obtained from STRAIGHT, and were corrected for the effects of the first two formant frequencies using the formula in [10]. Similarly,  $A_3^*$  were calculated from the speech spectrum using the formant frequencies obtained from Snack and were also corrected for the effects of the first two formants.  $A_3^*$  was additionally corrected for the effects of  $F_3$ . All the measures were calculated using the “VoiceSauce” software [16].

For each classification experiment, 70% of the utterances were selected randomly for training and the remaining 30% of utterances were used for testing. Utterances from a particular

speaker were used either for training or testing. Five experiments were conducted for each combination of acoustic measures and average accuracies were recorded. Note that for each utterance acoustic measures were calculated frame by frame and then averaged over the utterance.

In this paper, classification was done using an SVM classifier with a Radial Basis Function kernel. The LIBSVM toolkit [17] was used for training and testing. The results of the SVM classifier were compared with traditional MFCC features, using the first 12 MFCCs extracted from the vowel segment in each utterance. Due to the small number of vowels, training was implemented using 2 GMMs, each with 6 mixtures.

## 4. Analysis

$F_0$  and formant frequency values, averaged across all subjects for each age group, are provided in Table 2, and the means and standard deviations of  $F_0$  for each group are shown in Figure 1. Results are similar to [6, 8]. It is observed that  $F_0$  for male and female speakers is not distinguishable for the age groups 8–9 and 10–11. The  $F_0$  difference between male and female speakers becomes significant beginning from age 12, partly due to the drop in  $F_0$  for male speakers between age 12 and 15 [6].

Table 2: Mean and standard deviation (in parentheses) of fundamental frequency and formant frequency values for male and female speakers for each age group (in Hz)

Age group	Gender	$F_0$	$F_1$	$F_2$	$F_3$
8–9	female	267(40)	609(257)	2154(618)	3196(419)
	male	257(41)	578(236)	2109(620)	3170(428)
10–11	female	255(43)	629(242)	2242(532)	3170(411)
	male	253(41)	578(226)	2145(575)	3143(417)
12–13	female	239(33)	590(223)	2233(493)	3198(339)
	male	212(47)	546(207)	2093(469)	3053(383)
14–15	female	227(27)	594(198)	2113(409)	3002(334)
	male	150(45)	527(191)	2013(396)	2883(339)
16–17	female	223(25)	581(199)	2112(446)	3007(299)
	male	128(31)	490(193)	1952(361)	2804(347)

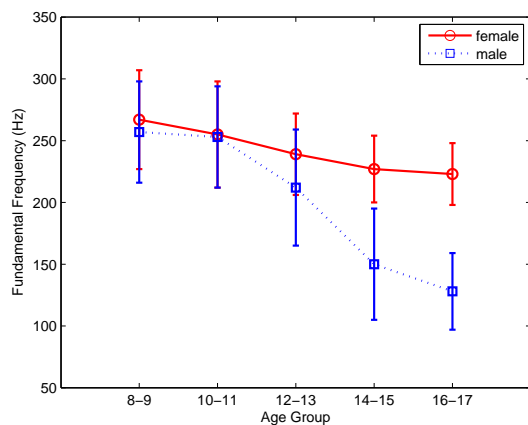


Figure 1:  $F_0$  averaged across all subjects is shown for each age group

Values of  $CPP$ ,  $HNR$  and  $H_2^* - H_4^*$ , averaged across all subjects for each age group, are provided in Table 3. The means and

standard deviations of  $CPP$  are shown in Figure 2. It can be seen from the figure that the difference in  $CPP$  between male and female speakers is not significant for age group 8–9, which is relatively of the same scale as the difference in  $F_0$ . For age groups 10–11 and 12–13, however, the difference between male and female speakers in  $CPP$  increases, which is relatively larger than the difference in  $F_0$ . With increasing age,  $HNR$  and  $H_2^* - H_4^*$  begin to differentiate male and female speakers from age 12 and 14, respectively; but the differences are overshadowed by the large standard deviations. This suggests that, the involvement of voice source measures, such as  $CPP$ , could improve gender classification accuracy for pre-adolescents, whereas  $F_0$  values do not help differentiate between male and female children’s speech.

Table 3: Mean and standard deviation (in parentheses) of  $CPP$ ,  $HNR$  and  $H_2^* - H_4^*$  values for male and female speakers for each age group (in dB)

Age group	Gender	$CPP$	$HNR$	$H_2^* - H_4^*$
8–9	female	22.42(3.00)	31.30(8.71)	2.61(7.80)
	male	22.59(3.04)	31.24(8.04)	3.09(7.52)
10–11	female	22.62(2.51)	31.25(7.84)	3.18(7.18)
	male	23.25(2.75)	30.14(7.39)	3.32(7.05)
12–13	female	22.97(2.28)	30.79(7.10)	2.96(6.59)
	male	23.88(2.93)	28.46(7.35)	3.90(6.47)
14–15	female	23.75(1.84)	31.57(7.47)	3.07(5.09)
	male	24.27(3.38)	25.06(8.05)	6.22(7.03)
16–17	female	23.21(2.25)	32.68(8.49)	2.41(5.49)
	male	24.68(2.40)	25.32(8.86)	6.21(5.86)

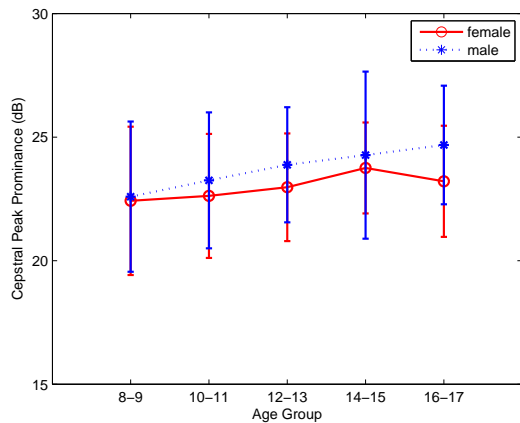


Figure 2: Cepstral Peak Prominence value averaged across all subjects is shown for each age group

## 5. Classification results

In this section, M0 is used to denote a feature set representing formant information ( $F_1, F_2, F_3, B_1$  and  $B_2$ ) and  $F_0$ , as in [1], and it is used as the baseline feature set. M1 is used to denote the feature set M0 with  $H_1^* - H_2^*$  and  $H_1^* - A_3^*$  from [1] which gave the best performance in that study. M2 denotes the feature set M0 with  $CPP$ . M3 denotes the feature set M0 with  $CPP$  and  $HNR$ . M4 denotes the feature set M0 with  $CPP, HNR$  and  $H_2^* - H_4^*$ . Table 4 summarizes these sets.

Table 4: Measure sets (M0-M4) used in the gender classification tests.

Set	M0	$H_1^* - H_2^*$	$H_1^* - A_3^*$	$CPP$	$HNR$	$H_2^* - H_4^*$
M0	✓					
M1	✓	✓	✓			
M2	✓			✓		
M3	✓			✓	✓	
M4	✓			✓	✓	✓

### 5.1. Results using additional voice-source related measures for each age group

Figure 3 compares gender classification accuracies from different measure sets. It can be seen from the figure that the classification accuracies of M4 are higher than the baseline and M1. Table 5 shows classification accuracies for each age group compared with the results for M0, M1 and also for the MFCC/GMM method.

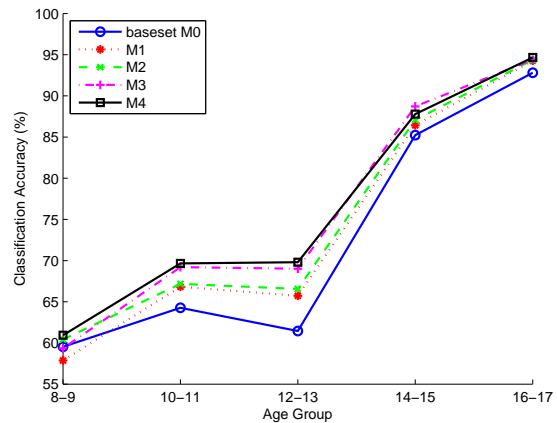


Figure 3: Gender classification accuracy for each age group using the measures sets M0, M1, M2, M3, M4.

Table 5: Gender classification accuracy for the different measurement sets (M0-M4) on each age group (in %). Boldface represents the highest accuracy for each age group

Age group	M0	M1	M2	M3	M4	MFCC/GMM
8–9	59.54	57.87	60.43	59.35	<b>60.93</b>	59.30
10–11	64.27	66.82	67.17	69.21	<b>69.66</b>	60.62
12–13	61.45	65.73	66.56	69.02	<b>69.81</b>	68.08
14–15	85.23	86.43	87.10	<b>88.71</b>	87.78	82.30
16–17	92.80	94.26	94.37	94.38	<b>94.66</b>	90.79

It can be seen that the addition of voice source measures  $CPP, HNR$ , and  $H_2^* - H_4^*$  constantly improved classification accuracies, compared to M0 and M1, for all age groups. An average of 3.2% improvement was achieved by adding measure  $CPP$  to the baseline set M0 (the M2 set). With the exception of age group 8–9, classification accuracies were further improved by adding the  $HNR$ . The change in classification accuracies by adding measure  $H_2^* - H_4^*$  was not significant (the M4 set). The performance of voice source measures set M4 is about 4.4% higher than the result for M0 and about 3% higher than the result for M1. A large improvement of about 8% is obtained on age group 12–13 when comparing M4 with M0.

Table 6: Gender classification accuracy for M4 set on each age group, distinguishing between males and females (in %).

Age group	8-9	10-11	12-13	14-15	16-17
M	56.66	69.37	67.47	87.30	93.38
F	65.32	69.88	72.13	88.19	95.95

Table 6 shows the classification accuracies of M4 set for males and females respectively. Interestingly, the accuracy is higher for females than that of males for all age groups. Similar results were reported in [1].

## 5.2. Discussion

The results in Table 5 show that using *CPP* and *HNR* is useful in improving gender classification accuracy for children’s speech. This suggests that the voice source measures *CPP* and *HNR* contain characteristics which are unique for young male and female speakers. Since *CPP* is highly correlated with breathiness [2], the results confirm that, in general, females are breathier than males [11]. Interestingly, *HNR* is higher for females than males for all age groups. The difference in *HNR* between females and males increases with increasing age. This is inconsistent with the expectation that females should have lower *HNR* values than males, since in general females are more breathy than males [11]. This result requires further exploration on what signal property contributed to the high *HNR* of females. As stated in [3]: “All kinds of signal properties may result in a noise-like appearance of the spectrum, such as a perturbation of the excitation signal (jitter and shimmer), rapid directional changes in fundamental frequency, formant transitions, and so forth.” A possible explanation for these results could be the interaction effects of the noise level perception. A recent study [18] showed that listener’s perception of noise levels in voice depends on the shape of the harmonic spectrum; but the interaction effects of voice quality perception are not well understood.

The measure  $H_2^* - H_4^*$  also assisted in improving the classification accuracies, with the exception of age group 14–15, suggesting that the mid-frequency tilt also differentiates between the male and female speech spectra.

A large improvement of about 8% is obtained for the age group 12-13 when comparing M4 with M0. This could be attributed to the fact that puberty of males and females begins at around 11 [6]. Adding other measures, such as formant amplitudes, didn’t improve the classification accuracy significantly.

For age group 8–9, the classification accuracy for all measure sets are below 61%. The improvement by adding features *CPP*, *HNR* and  $H_2^* - H_4^*$  is not significant.

Considering the performance of all age groups, the addition of measures *CPP*, *HNR* and  $H_2^* - H_4^*$  improved classification accuracy by 4.4% compared with the baseline feature set. When compared with M1, the feature set M4 provides about 3% improvement (on average) for all age groups. While the performances of feature set M4 are similar to the MFCC/GMM results for age groups 8–9 and 12–13, the classification accuracies for M4 is about 9%, 5% and 4% higher for age groups 10–11, 14–15 and 16–17, respectively.

## 6. Summary and conclusions

In this paper, we applied measures related to the voice source in gender classification using children’s voices and compared the results with those in [1]. The experiments were done using

the CID database which consisted of 3418 utterances spoken by 174 male and 140 female speakers. Measures related to the voice source measures and vocal tract were extracted from 5 target vowels and applied in gender classification tests.

The feature set consisting of  $F_0$ , the first three formant frequencies ( $F_1$ ,  $F_2$  and  $F_3$ ) and the first two bandwidths ( $B_1$  and  $B_2$ ) were used as baseline feature set (M0). Features were added to the baseline set to test their effect on gender classification. Results show that adding the three measures *CPP*, *HNR* and  $H_2^* - H_4^*$  yielded best overall performance, suggesting that measures related to breathiness and mid-frequency tilt carry discriminative information for automatic gender classification. The accuracy improvements of adding voice source features were highest for age group 12–13. This could be attributed to the fact that puberty occurs at around age 11. After age 13, the accuracy improvements of adding voice source features decreased as the role of  $F_0$  became more prominent.

Future work will focus on applying voice source measures to continuous speech.

## 7. Acknowledgements

This work was supported in part by the NSF.

## 8. References

- [1] Y.-L. Shue and M. Iseli, “The role of voice source measures on automatic gender classification,” in *Proceedings of ICASSP*, 2008, pp. 4493–4496.
- [2] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, “Acoustic correlates of breathy vocal quality,” *Journal of Speech and Hearing Research*, vol. 37, pp. 769–778, 1994.
- [3] G. de Krom, “A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals,” *Journal of Speech and Hearing Research*, vol. 36, pp. 254–266, 1993.
- [4] K. Wu and D. G. Childers, “Gender recognition from speech. part i: coarse analysis,” *Journal of the Acoustical Society of America*, vol. 90, pp. 1828–1840, 1991.
- [5] P. Busby and G. Plant, “Formant frequency values of vowels produced by preadolescent boys and girls,” *Journal of the Acoustical Society of America*, vol. 97, pp. 2603–2606, 1995.
- [6] S. Lee, A. Potamianos, and S. Narayanan, “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters,” *Journal of the Acoustical Society of America*, vol. 105, pp. 1455–1468, 1999.
- [7] T. L. Perry, R. N. Ohde, and D. H. Ashmead, “The acoustic bases for gender identification from children’s voices,” *Journal of the Acoustical Society of America*, vol. 109(6), pp. 2988–2988, 2001.
- [8] M. Iseli, Y.-L. Shue, and A. Alwan, “Age, sex, and vowel dependencies of acoustic measures related to the voice source,” *Journal of the Acoustical Society of America*, vol. 121, pp. 2283–2295, 2007.
- [9] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. Guid, 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.
- [10] M. Iseli and A. Alwan, “An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation,” in *ICASSP*, vol. 1, 2004, pp. 669–672.
- [11] D. Klatt and L. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, 1990.
- [12] J. Miller, S. Lee, R. Uchanski, A. Heidebreder, B. Richman, and J. Tadlock, “Creation of two children’s speech databases,” in *Proceedings of ICASSP*, vol. 2, May 1996, pp. 849–852.
- [13] J. Kreiman, B. Gerratt, and N. A. nanzas Barroso, “Measures of the glottal source spectrum,” *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 595–610, 2007.
- [14] K. Sjölander, “Snack sound toolkit,” KTH Stockholm, Sweden, 2004, <http://www.speech.kth.se/snack/> (last viewed Aug. 2009).
- [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 of ICSLP*, 1998.
- [16] Y.-L. Shue, “VoiceSauce: a program for voice analysis,” 2010, <http://www.ee.ucla.edu/~spapl/voicesauce/>.
- [17] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (last viewed Aug. 2009).
- [18] B. Gerratt and J. Kreiman, “A spectral-slope compensated scale for measuring perception of vocal aperiodicity,” *Journal of the Acoustical Society of America*, vol. 127, pp. 2022–2022, 2010.