Comparison of Text-Dependent Method for Gender Identification

Sumit Kumar Banchhor, S.K.Dekate

Abstract— Differences of physiological properties of the glottis and the vocal track 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 gender classification. Acoustics measures of voice sources were extracted from 10 utterances spoken by 20 male and 20 female talkers (aged 19 to 25 year old). The difference of speech long term features, including zero crossing rate, short time energy, and spectrum flux between male and female is studied. The result shows that the estimation of short time energy reflects more effectively, the difference in male and female voice than zero crossing rate and spectrum flux.

Index Terms— gender classification gender identification, voice source.

I. INTRODUCTION

Gender-based differences in human speech are due 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 gender classification.

One of the most important factors which cause an increase in an intelligent agent's decision performance is to possess information about related pattern or subject, as much as possible. Value and usefulness of information, depends on its discrimination ability between critical states of decision making. Supplementary knowledge with more discrimination ability between critical states of decision-making leads to more increase in performance of agent's decision. Sex is among the most important information that discriminate between diverse speakers. As a practical example, if a speaker recognizer agent could recognize speaker sex, it can both restrict its search between enrolled speakers to decrease decision time and increase its recognition performance. In another application, agent may offer better suggestion to a remote customer in order to buy a company's new product (such as new titles of a publisher) by using its knowledge about his sex. Due to above reasons, speaker sex identification research have recently gained much attention in the field of automatic speech processing.

Manuscript received May 30, 2011.

Sumit Kumar Banchhor, Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University/Rungta College of Engineering and Technology/ Bhilai, India, 9893880318,
(E-Mail: sumit.9981437433@gamil.com).
S.K.Dekate, Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University/Rungta College of Engineering and Technology/ Bhilai, India. 9301771560, (E-Mail: sd.rungta.college@gmail.com).

II. PREVIOUS WORK

GMM and neural networks were trained for age interval and sex identifications using speech short term features[1]. Homayounpour and Khosravi make an effort on identifying speaker age interval using SVM model and MFCC and LPCC features[2]. Speaker age interval and sex identification were done using GMMs[3],[4].

It is well known that $F0$ values for male talkers drop during adolescence due to a lengthening and thickening of the vocal folds. $F0$ for adult males is typically around 120 Hz, while $F0$ for adult females is around 200 Hz. This effect is mostly due to a lengthening and thickening of the male vocal folds[5].

It is also well known that, due to vocal tract length differences, adult males exhibit lower formant frequencies than adult females[5]. Interestingly, for preadolescent children, studies also found lower formant frequencies for boys compared to girls of ages 5-6 [6], 7-8 years [7], and ages 5, 7, 9, and 11 years (for Australian English) [8]. These findings imply that, overall, boys have larger vocal tracts than girls. Statistical analysis of children speech confirmed that formant frequencies ($F1$, $F2$, $F3$), and not $F0$, differentiate gender for children as young as 4 years of age, while formant frequencies plus $F0$ differentiate gender after 12 years of age[9]. 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 $F0$. For adult speech, automatic gender classification has been presented, which used linear predictive coding (LPC)-derived measures that represent the vocal tract[10].

Other measures like $F0$, formant frequencies, and spectral envelope are presented as a function of age for talkers from 5 to 50 years old. For $F0$, the study showed a drop between ages 12 and 15 for males and a drop of $F0$ variation for all talkers between ages 5 and 15. Formant frequencies ($F1$, $F2$, $F3$) 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[11].

We analyzed age, sex, and vowel dependencies, for talkers between the ages of 8 and 39, of the following three voice source measures: $F0$[12]; $H*1 - H*2$, the difference of the first two source spectral harmonic magnitudes (related to the open quotient1 [13]); 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 [13]).

The asterisk indicates a correction for the influence of vocal tract resonances [14]. 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 $F0$, $H*1 - H*2$, and $H*1 - A*3$ for adult female talkers compared to adult male talkers[15].

In this paper, acoustics measures of voice sources were extracted from 10 utterances spoken by 20 male and 20 female talkers (aged 19 to 25 year old). The difference of speech long term features, including zero crossing rate, short time energy, and spectrum flux between male and female is studied.

III. SPEECH DATA BASE FORMULATON

Speech recording from age group, 19-25 were taken. Each recording was of the form "Now this time you go", where the target vowel 'V' was /th/. This utterance where spoke at the habitual speaking level and most talkers repeated the phrases 10 times. For the analysis, only the manually segmented target vowels where used.

IV. METHODOLOGY

The target vowel was manually segmented using GOLDWAVE software and stored with .wav extension.

V. EXPERIMENT AND RESULTS

A. RESULT USING ZERO CROSSING RATE

It indicates the frequency of signal amplitude sign change. To some extent, it indicates the average signal frequency as:

$$ZCR = \frac{\sum_{n=1}^{N} |\text{sgn } x(n) - \text{sgn } x(n-1)|}{2N}$$

Where *sgn[]* is a signum function and *x(m)* is the discrete audio signal.

In mathematical terms, a "zero-crossing" is a point where the sign of a function changes (e.g. from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function. The zero-crossing is important for systems which send digital data over AC circuits, such as modems, X10 home automation control systems, and Digital Command Control type systems for Lionel and other AC model trains. Counting zero-crossings is also a method used in speech processing to estimate the fundamental frequency of speech.

Zero crossing rate is important because they abstract valuable information about the speech and they are simple to compute.

Table I. Average Zero Crossing Rate Of Male And Female Voice

Frames / Sex	5	10	15	20	25	30	35	40
Male	8.6	9.75	8.8	8	8.85	7.3	8.45	7.85
Female	11.8	10.1	9.95	10.4	9.3	8.3	8.85	8.2

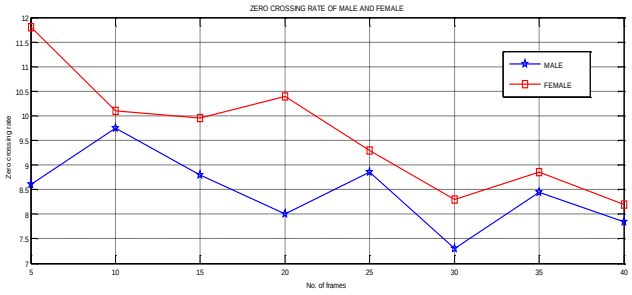

Figure 1. Zero crossing rate for male and female voice

Figure 1 displays the zero crossing rate (ZCR) of male and female. It shows that ZCR for female is little higher than male.

B. RESULT USING SHORT TIME ENERGY

The short time energy measurement of a speech signal can be used to determine voiced vs. unvoiced speech. Short time energy can also be used to detect the transition from unvoiced to voiced speech and vice versa. The energy of voiced speech is much greater than the energy of unvoiced speech.

We record the i/p signal at fs=8KHz. Now using Hamming window with the following specifications : Window size=256 samples, Window step=100 samples, Window overlap=156 samples and number of frames = (length of i/p – window size)/(window step), we calculate the STE for each frame using the following formula.

$$E = \sum_{m=0}^{N-1} \left| x(n)^2 / (N) \right|$$

TABLE II. short time energy OF MALE AND FEMALE VOICE

Frames / Sex	5	10	15	20	25	30	35	40
Male	0.05	0.08	0.09	0.09	0.1	0.08	0.07	0.06
Female	0.92	1.36	1.5	1.38	1.17	0.88	1.06	0.73

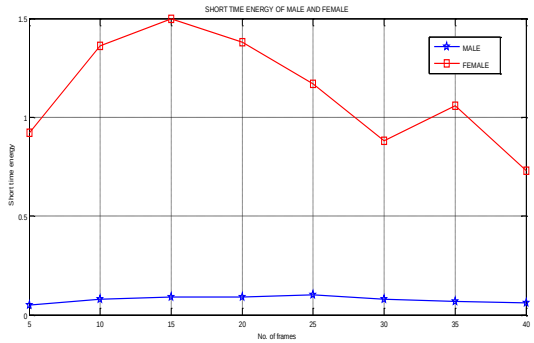

Figure 2. Short time energy for male and female voice

Figure 2 displays the short time energy (STE) of male and female. It shows that STE for female are much higher than male.

C. RESULT USING SPECTRUM FLUX

Spectral flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. More precisely, it is usually calculated as the 2-norm (also known as the Euclidean distance) between the two normalized spectra. Calculated this way, the spectral flux is not dependent upon overall power (since the spectra are normalized), nor on phase considerations (since only the magnitudes are compared). If there is a transient or a sudden attack, the change in energy will be denoted by a jump in the difference of energy between consequent frames. It is important to note that after taking the difference in the spectrums, a positive difference value indicates a rise in energy while a negative difference value indicates a dip in energy. If this method is employed to detect transients, a threshold value should be set only for a positive difference value.

We record the i/p signal at fs=8KHz. Now using Hamming window with the following specifications : Window size=256 samples, Window step=100 samples, Window overlap=156 samples and number of frames = (length of i/p – window size)/(window step), we calculate the STE for each frame using the following formula.

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log A(n,k) - \log A(n-1,k)]^2$$

Where *A(n,k)* is the discrete Fourier transform of the nth frame of input signal.

$$A(n,k) = \sum_{m=\infty}^{\infty} x(m)w(nL-m)e^{j2\Pi km/L}$$

Where L is the window length, k is the order of DFT, N is the total number of frames.

Table iii. Spectrum flux of male and female voice

Sex \ Frames	5	10	15	20	25	30	35
Male	0.24	0.16	0.56	1.9	0.35	0.26	0.45
Female	0.07	0.06	0.35	0.87	0.065	0.06	0.03

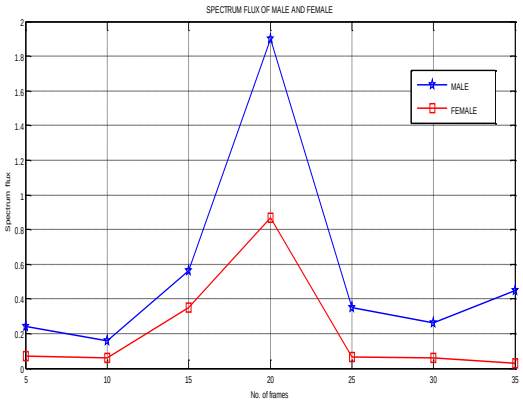

Figure 3. Spectrum flux for male and female voice

Figure 3 displays the spectrum flux (SF) of male and female. It shows that SF for male are much higher than female.

VI. DISCUSSION AND CONCLUSION

TABLE IV. comparison of various parameters

SL.NO.	PARAMETER	MALE	FEMALE
1	Zero crossing rate	LESS	GREATER
2	Short time energy	<= 0.1	>= 0.88
3	Spectrum flux	1.9	0.87

In this paper, we examined the role of voice source measure in gender identification and compared the results to perceptual experiments performed on the same database. Voice source measures were extracted from a large database of utterance spoken by 20 males and 20 females.

We used three different parameters in the analysis. From the experiments, we could observe evident results for zero crossing rate, short time energy, and spectrum flux. Zero crossing rate, and short time energy of female are larger than those of male. Zero crossing rate of female are little larger than those of male whereas short time energy of female are much larger than those of male. Spectrum fluxes of male are larger than those of female.

The result shows that the estimation of short time energy reflects more effectively the difference in male and female voice than zero crossing rate and spectrum flux.

These coefficients contain useful information about speakers gender classification and employing them in such a process lead to decrease in gender identification error rate.

REFERENCES

1. M. M. Homayounpour, B. Mobarakabadi, N. Hamidi, “Age interval and sex identification based on voice, using GMM and neural networks”, 11th Iranian Conference on Electrical Engineering, pp 304-311, 2003, (in persian).
2. M. M., Homayounpour, M. H., Khosravi,., "Age Identification Using Support Vector Machine", IKT2003, pp. 615-622, 2003.
3. Nobuaki Minematsu, Keita Yamauchi, and Keikichi Hirose, "Automatic estimation of perceptual age using speaker modeling techniques", 8th European Conference on Speech Communication and Technology, EUROSPEECH, 2003.
4. Peder A. Olsen, Satya Dharanipragada, "An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models", 8th European Conference on Speech Communication and Technology, EUROSPEECH, 2003.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.

10. 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.
11. 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.
12. 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.
13. 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.
14. 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, vol. 1, pp. 669–672, May 2004.
15. 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.
16. Ruan boyao. The application of PCNN on speaker recognition based on spectrogram. Master Degree Dissertations of Wuyi University. 2008.
17. An expert spectrogram reader: A knowledge-based approach to speech recognition Zue, V.; Lamel, L.; Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86. Volume: 11, pp. 1197 – 1200, 1986
18. Mathew J.Paiakal and Michael J.Zoran. Feature Extraction form Speech Spectrogram Using Multi-Layered Network Models. Tools for Artificial Intelligence, 1989. Architectures, Languages and Algorithms, IEEE International Workshop on Volume, Issue, 23-25, pp. 224 – 230, Oct 1989.
19. Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigne. Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication. Volume 27, Issue 3-4, pp. 187 – 207, Apr 1999.
20. Yang yang. Voiceprint Recognition Technology and Its Application in Forensic Expertise. Master Degree Dissertations of Xiamen University. 2007.
21. V.W. Zue and R.A. Cole, "Experiments on Spectrogram Reading, "*IEEE Conference Proceedingo, ICASSP*, Washington D.C.,1979, pp. 116-119

AUTHORS PROFILE

Sumit Kumar Banchhor, is currently pursuing masters degree program in digital electronics in Chhattisgarh Swami Vivekananda Technical University, India,
PH:+91 9893880318.E-mail:sumit.9981437433@gmail.com

S. K. Dekate, is Associate Professor in Rungta College of Engineering and Technology in Chhattisgarh Swami Vivekananda Technical University, India,
PH:+91 9301771560. E-mail: sd.rungta.college@gmail.com