Estimating the Stability and Dispersion of the Biometric Glottal Fingerprint in Continuous Speech

P. Gómez, A. Álvarez, L. M. Mazaira, R. Fernández, V. Rodellar

Grupo de Informática Aplicada al Procesado de Señal e Imagen
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, Boadilla del Monte
E-28660 Madrid, Spain
pedro@pino.datsi.fi.upm.es

Abstract

The speaker's biometric voice fingerprint may be derived from voice as a whole, or from the vocal tract and glottal signals, after separation by inverse filtering. This last approach has been used by the authors in early work, where it has been shown that the biometric fingerprint obtained from the glottal source or related speech residuals gives a good description of the speaker's identity and meta-information, as gender or age. In the present work a new technique is proposed based on the accurate estimation of the glottal residual by adaptive removal of the vocal tract, and the detection of the glottal spectral singularities in continuous speech. Results on a reduced database of speakers demonstrate that the biometric fingerprint estimation is robust, and shows low intra-speaker variability, which makes it a useful tool for speaker identification as well as for pathology detection, and other fields related with speech characterization.

1. Introduction

In previous work it has been shown that the biometric fingerprint obtained from the glottal source or related speech residuals after careful removal of the vocal tract function by inverse filtering [1][2] gives a good description of the speaker's identity and meta-information, as gender. The main inconvenience in using this technique was the requirement of using phonation cycles for the estimation of the fingerprint parameters, and its frame-based stationary nature. Besides, the variability of the estimates was strongly conditioned by the glottal gesture (tension, pitch extent and radiation – chest, mouth or head). To solve these problems a new technique is proposed, based on the accurate estimation of the glottal residual by adaptive removal of the vocal tract, and the detection of the glottal spectral singularities using lateral inhibition. The paper is organized as follows: an overview of the adaptive estimation of the glottal source and the vocal tract is briefly summarized. The glottal residual (glottal pulse, or first derivative of the glottal source) and the glottal source power spectral densities obtained by FFT in succesive sliding windows are scanned to detect their spectral singularities (maxima and minima) as these may be shown to be strongly related to vocal fold biomechanics [5]. Estimates from male and female voice show that these singularities are gender-specific. Based on the estimations from a wide data base of 100 subjects, it may be shown that specific parameters present in the glottal fingerprint may be used for gender classification [7]. The extension of the present results for age and gender characterization following well established research based on formant estimates [11] is foreseen.

2. Glottal Source adaptive estimation

The key for the accurate estimation of the glottal source is to obtain a good estimation of the vocal tract transfer function, and vice-versa. Traditionally it has been considered that the glottal source has a power spectral density of *1/f*. This assumption, being acceptable as far as its spectral envelope is concerned, hiddens the fact that the glottal signals have spectral signatures of their own, on which the fingerprint of vocal fold biomechanics can be found and used for specific applications, as is the biometrical description of the speaker or pathology detection [4].

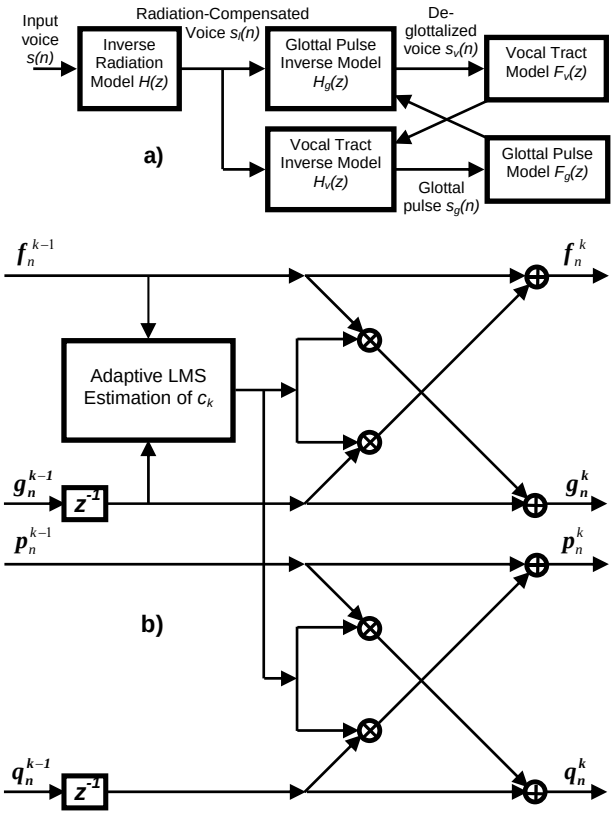

Figure 1. *a) Iterative estimation of the vocal tract transfer function $F_v(z)$ and the glottal pulse residual $s_g(n)$. b) Paired adaptive-fixed lattice section to implement parallel function estimation and removal.*

The iteration is based on the successive application of the following loop as shown in Figure 1.a:

- Estimate the inverse glottal pulse model $F_g(z)$ from input voice using an *order-2* adaptive lattice (upper lattice in Figure 1.b).
- Remove the glottal pulse from input voice using a paired fixed lattice (lower lattice in Figure 1.b) fed with the reflection coefficients obtained by the adaptive lattice for the glottal function. The resulting trace $s_v(n)$ will keep the vocal tract information but the glottal information will be diminished.
- Estimate the vocal tract transfer function $F_v(z)$ from this last trace using another adaptive lattice (typically of order *20-30*), similar to Figure 1.b.
- Remove the vocal tract transfer function from input voice using a fixed lattice (lower lattice in Figure 1.b) fed with the reflection coefficients obtained by the adaptive lattice for the vocal tract function. The resulting trace $s_g(n)$ will keep the glottal information but the vocal tract influence will be small.

This iteration is repeated several times till successive estimations of the vocal tract transfer function is almost free from glottal source information, and vice-versa. The results produced for the present paper were obtained after an initialization lap and two more iterations. The details for the adaptive lattice implemented may be found in [8]. An example of the traces obtained from the utterance of vowel /a/ by a typical male speaker is shown in Figure 2.

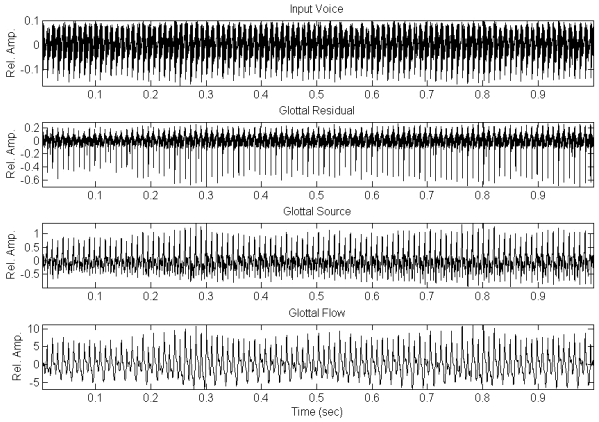

Figure 2. *A typical male speaker utterance of vowel /a/ and derived glottal traces. From top to bottom: input voice, glottal residual after adaptive inverse removal of the vocal tract, glottal source, and glottal flow*

3. Estimating the biometric fingerprint

The fingerprint is estimated from the FFT power spectral density of the glottal residual, after normalization to obtain the positions of envelope singularities as follows:

- The glottal residual and glottal source are windowed in *512* sample frames sliding *2 msec* and the power spectral density of each window is estimated by FFT in logarithmic (dB) scale as shown in Figure 3 (full line) for two examples of typical male and female speakers.
- The envelopes of the power spectral densities of these short-time power spectra are estimated (dot line).
- The maxima (*) and minima (◊) found on the respective envelopes are detected and their amplitudes and frequencies collected as two lists of ordered pairs: $\{T_{Mk}, f_{Mk}\}$ and $\{T_{mk}, f_{mk}\}$, with k the ordering index.
- The largest of all maxima (T_{Mm}, f_{Mm}) is used as a normalization reference both in amplitude and in frequency as given by:

$$\left.\begin{aligned}\tau_{Mk} &= T_{Mk} - T_{Mm}\\ \tau_{mk} &= T_{mk} - T_{Mm}\end{aligned}\right\};\quad 1 \le k \le K \qquad (1)$$

$$\left.\begin{aligned}\varphi_{Mk} &= \frac{f_{Mk}}{f_{Mm}}\\ \varphi_{mk} &= \frac{f_{mk}}{f_{Mm}}\end{aligned}\right\};\quad 1 \le k \le K \qquad (2)$$

An important parameter derived from the ordered minima and maxima pairs is the *slenderness* factor, defined on each "V" trough formed by each minimum and the two neighbor maxima, which may be defined as:

$$\sigma_{mk} = \frac{f_{Mm}\left(2T_{mk} - T_{Mk+1} - T_{Mk}\right)}{2\left(f_{Mk+1} - f_{Mk}\right)};\quad 1 \le k \le K \qquad (3)$$

The results of estimating the normalized singularity parameters for the cases under study may be seen in Figure 4.

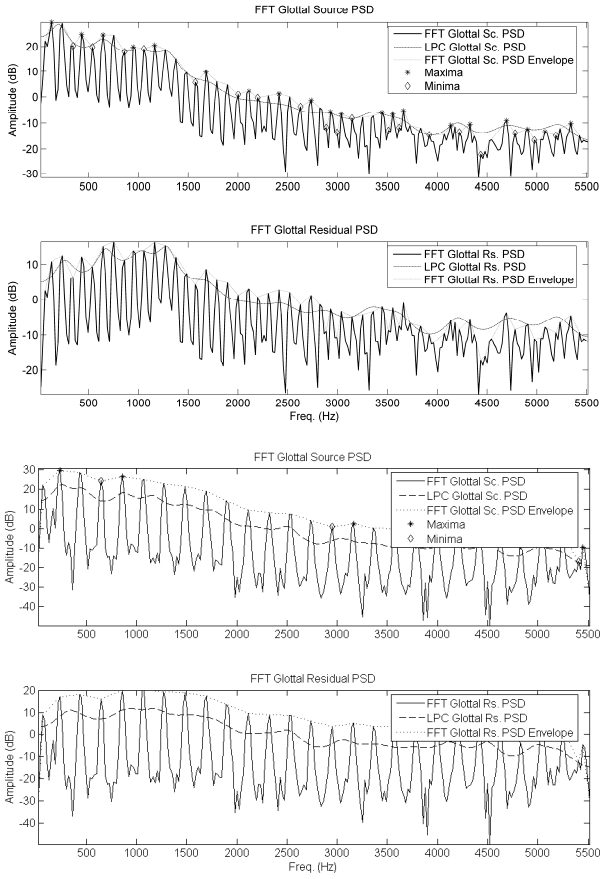

Figure 3. *Short-term Power Spectral Density of Glottal Signals. Two top plots: Glottal Source(showing superimposed the singularities) and Glottal Residual for a typical male speaker (vowel /a/). Two bottom plots: idem for a typical female speaker (vowel /a/).*

4. Materials and methods

To establish the validity of the proposed biometric fingerpring data recorded on a population set of 100 normal speakers equally distributed by gender were used. Subject ages ranged from 19 to 39, with an average of 26.77 years and a standard deviation of 5.75 years. The normal phonation condition of speakers was determined by electroglottographic, video-endoscopic and GRBAS [9] evaluations. The recordings at a sampling rate of 44,100 Hz consisted in three utterances of the vowel */a/* produced in different sessions of about *3 sec* per record, a *0.2 sec* segment derived from the central part being used in the experiments. For presentation purposes the traces were re-sampled at 11,025 Hz. This database was fully parameterized to obtain the singularity biometric fingerprint described in section 3. The normalized biometric fingerprints for the speakers presented are given in Figure 4.

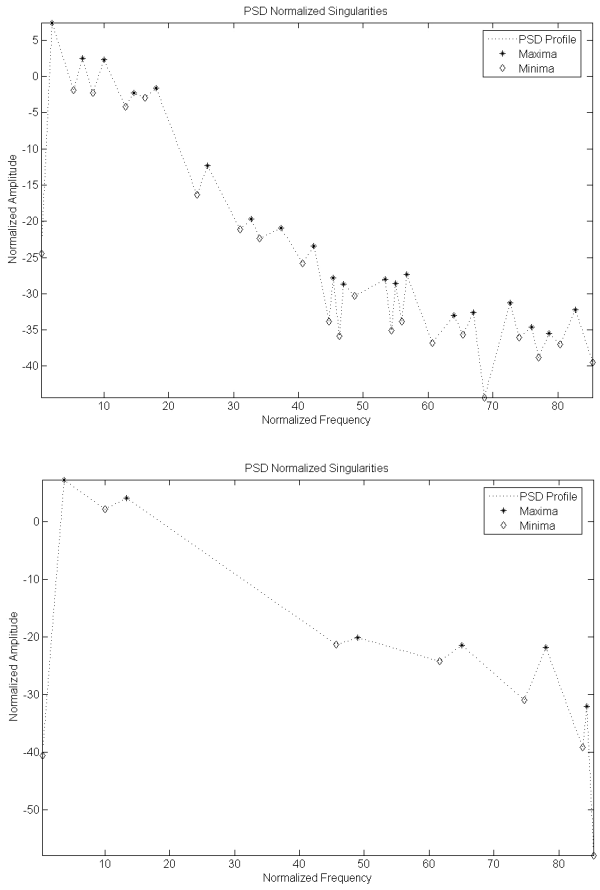

Figure 4. *Normalized singularity profiles for the male (top and female (bottom) typical utterances of vowel /a/*

A first inspection of the fingerprints shows that the one from the male speaker exhibits deeper "V" troughs at lower relative frequencies than the female case. This is consistent with the biomechanical explanation of the nature of peaks and troughs, as these are based on the mechanical resonances and anti-resonances of the systems of masses and springs describing vocal fold vibration. In general, female vocal folds show more stiff links among body and cover masses, and this would explain the lower amount of less sharp anti-resonances (see [3] and [5] for a wider explanation).

Important considerations are robustness and intra-speaker variability of estimates. The processing of the *0.2 sec* segments in *512 sample* windows sliding in *2 msec* steps produce around *76* estimates per segment. The positions of the 3 minima and maxima for the male and female speakers (plus the origin value) vs time are given in Figure 5.

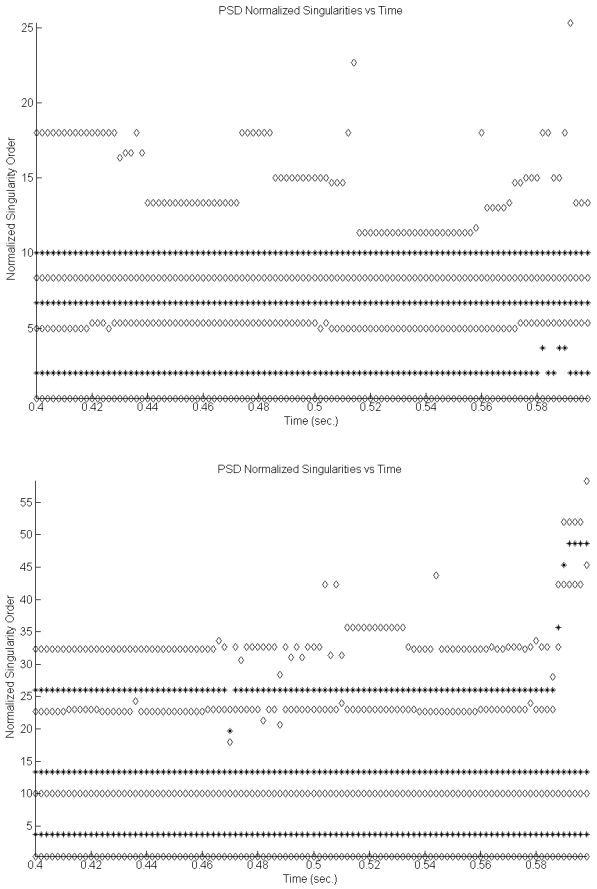

Figure 5. *Relative positions of the first 7 singularity points and the origin for a 0.2 sec segment of the reference male and female traces*

It may be observed that lower order are more stable than higher order estimates. This finding is consistent with the biomechanical explanation of the estimates. Lower frequency troughs and peaks are due to larger vocal fold masses, which for a given articulation and load do not change substantially during phonation, whereas higher order singularities are due to irregular small mass distributions on the cord, which may suffer important alterations during phonation and are more sensitive to vocal tract coupling effects. This is especially so when analyzing results from complete sentences including different voiced sounds, not shown here for the lack of extent.

5. Results and discussion

The issue of intra-speaker variability may be better illustrated giving the statistical dispersion of the estimates for sustained vowels as the cases studied in normalized amplitude and frequency as given in Figure 5.

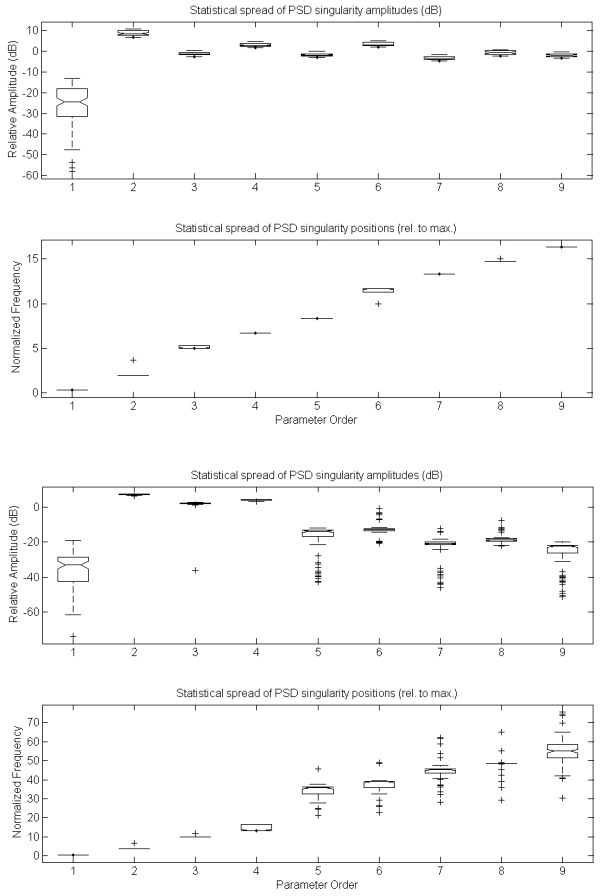

Figure 6. *Statistical distribution of the first 8 singularity points and the origin for a 0.2 sec segment of the traces shown in Figure 5. From top to bottom: Amplitudes (male), Normalized Frequencies (male), Amplitues (female), and Normalized Frequencies (female)*

The results shown confirm that the singularities in the male case are found at lower frequencies and that low frequency estimates are more stable than high frequency ones. This also explains why the spread of the estimates seems to be a little larger in the female case, as they appear at higher frequencies due to the more tense nature of female voice. As a general conclusion, it may be said that the intra-speaker variability of the low order singularities is small in sustained vowel-like utterances for normophonic speakers. This study has to be carried out to pathological cases as well and could be used for pathology detection. The estimations from the typical female voice shown are a little bit less stable. In previous work [6] it was shown that estimates from the glottal source may be used in the determination of the biomechanical parameters of the fold body, whereas the biomechanics of the fold cover could be obtained from the power spectral density of the mucosal wave correlate. An important pending study is the use of the glottal source or the glottal residual in determining the specific speaker's glottal profile.

6. Conclusions

The work shown is a generalization of prior studies using non-adaptive estimations of the vocal tract on short segments of vowels. The use of adaptive estimations allow a better accuracy in the estimates of the vocal tract, and consequently on the glottal signals. The extension of the glottal spectra singularities to time-varying conditions allow a better description of the non-stationary processes appearing in vocal fold vibration even in the production of sustained sounds. This improvement in the estimates may help in conducting more careful studies about inter-speaker and intra-speaker variability to extend the use of the glottal source spectral fingerprint to speaker identification and characterization applications [10][11].

7. Acknowledgments

This work is funded by grants TIC2003-08756, TEC2006-12887-C02-00 from Plan Nacional de I+D+i, Ministry of Education and Science, and project HESPERIA from the Program CENIT, CDTI, Ministry of Industry, Spain.

8. References

[1] Alku, P., "An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulseform", *Proc. of the ICASSP'92*, pp. II/29-32.

[2] Akande, O. O. and Murphy, P. J., "Estimation of the vocal tract transfer function with application to glottal wave analysis", *Speech Communication*, Vol. 46, No. 1, May 2005, pp. 1-13.

[3] Berry, D. A., "Mechanisms of modal and non-modal phonation", *J. Phonetics*, Vol. 29, 2001, pp. 431-450.

[4] Godino, J. I., Gomez, P., Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Trans Biomed. Eng*. Vol. 51, 2004, pp. 380-384.

[5] Gómez, P., Godino, J. I., Díaz, F., Álvarez, A., Martínez, R., Rodellar, V., "Biomechanical Parameter Fingerprint in the Mucosal Wave Power Spectral Density", *Proc. of the ICSLP'04, 2004*, pp. 842-845.

[6] Gómez, P., Martínez, R., Díaz, F., Lázaro, C., Álvarez, A., Rodellar, V., Nieto, V., "Estimation of vocal cord biomechanical parameters by non-linear inverse filtering of voice", *Proc. of the 3rd Int. Conf. on Non-Linear Speech Processing NOLISP'05*, Barcelona, Spain, April 19-22 2005, pp. 174-183.

[7] Gómez, P., Rodellar, V., Álvarez, A., Lázaro, J. C., Murphy, K., Díaz, F., Fernández, R., "Biometrical Speaker Description from Vocal Cord Parameterization", *Proc. of ICASSP'06*, Toulouse, France, 2006, pp. 1036-1039.

[8] Haykin, S., *Adaptive Filter Theory*, (4th Ed.), Prentice-Hall, Upper Saddle River, NJ, 2001.

[9] Hirano, M., Hibi, S., Yoshida, T., Hirade, Y., Kasuya, H., and Kikuchi, Y., "Acoustic analysis of pathological voice. Some results of clinical application," *Acta Otolaryngologica*, vol. 105, no. 5-6, pp. 432-438, 1988.

[10] Nickel, R. M., "Automatic Speech Character Identification", *IEEE Circuits and Systems Magazine*, Vol. 6, No. 4, 2006, pp. 8-29.

[11] Whiteside, S. P., "Sex-specific fundamental and formant frequency patterns in a cross-sectional study," *J. Acoust. Soc. Am.*, vol. 110, no. 1, pp. 464–478, 2001.