A HYBRID PARAMETERIZATION TECHNIQUE FOR SPEAKER IDENTIFICATION

P. Gómez, A. Álvarez, L. M. Mazaira, R. Fernández, V. Nieto, R. Martínez, C. Muñoz, V. Rodellar

GIAPSI, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n
28660 Boadilla del Monte, Madrid, Spain
phone: + (34) 913367384, fax: + (34) 913366601, e-mail: pedro@pino.datsi.fi.upm.es
web: www.mapaci.com

ABSTRACT

Classical parameterization techniques for Speaker Identification use the codification of the power spectral density of raw speech, not discriminating between articulatory features produced by vocal tract dynamics (acoustic-phonetics) from glottal source biometry. Through the present paper a study is conducted to separate voicing fragments of speech into vocal and glottal components, dominated respectively by the vocal tract transfer function estimated adaptively to track the acoustic-phonetic sequence of the message, and by the glottal characteristics of the speaker and the phonation gesture. The separation methodology is based in Joint Process Estimation under the uncorrelation hypothesis between vocal and glottal spectral distributions. Its application on voiced speech is presented in the time and frequency domains. The parameterization methodology is also described. Speaker Identification experiments conducted on 245 speakers are shown comparing different parameterization strategies. The results confirm the better performance of decoupled parameterization compared against approaches based on plain speech parameterization.

1. INTRODUCTION

Traditionally the idea that the glottal source does not contribute to the characterization of voice, as it is described by a transfer function of *1/f* has been taken as granted. Nevertheless recent works have shown that the glottal signals convey interesting features both in time and frequency which may be used for the speaker's characterization, speaker identification and pathology detection, among others [1]-[5]. The work of Plumpe et al. [6] is especially relevant, suggesting the use of parameters obtained from the time-domain glottal source description in speaker identification experiments. In preliminary work [7] it has been shown that the glottal source conveys important biometric information [8], which may be used in speaker's identification and verification tasks. The aim of the present work is to determine a parameterization technique taking into account the spectral characteristics of voicing speech to be decomposed into vocal tract and glottal source parameter templates to be used in speaker identification for security and forensic applications. Section 2 will explain the parameterization principles, section 3 will describe the experimentation framework, and the results will be presented and discussed in section 4, conclusions being exposed in section 5.

2. ESTIMATING THE GLOTTAL SOURCE

The speech source-filter model shown in Figure 1 assumes that voiced speech is generated by a glottal excitation $u_g(n)$ spectrally transformed by the vocal tract with transfer function $F_v(z)$ to produce the speech signal before radiation $s_l(n)$.

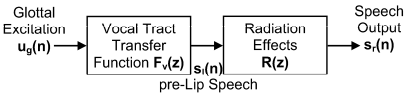

Figure 1. Generation model for voiced speech.

Many methods are available to cancel the influence of the vocal tract to estimate the glottal source [9][10], although not granting the statistical separation between the vocal tract impulse response and glottal excitation. The method proposed in the present approach grants that the estimates are orthogonal in terms of correlation, as described in Figure 2.

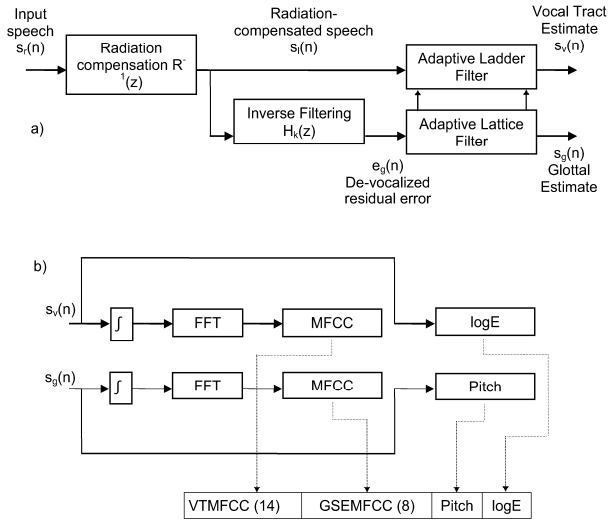

Figure 2. a) General framework to separate vocal from glottal characteristics by adaptive joint estimation. b) Parameterization scheme used in the experiments.

The separation technique is based on the inverse filtering of the voiced signal $s_l(n)$ to produce a residual $e_g(n)$ where the vocal tract has been removed by an *order-k* filtering process. The residual is used as the reference signal in an Adaptive Lattice-Ladder filter for Joint-Process Estimation [11] on the radiation-compensated speech $s_l(n)$. The main hypothesis

used in this approach is that $s_l(n)$ is produced by a glottal excitation $u_g(n)$ which may be seen as composed by a train of delta pulses $\delta_g(n)$ plus a glottal residual $u_r(n)$. The impulse response from the glottal closure produces the vocal component $s_v(n)$ (the impulse response of the vocal tract during the closed phase). The glottal component $s_g(n)$ results from the injection of flow $u_r(n)$ during the new open phase. Therefore the glottal excitation could be described as

$$u_g(n) = u_r(n) + \delta_g(n) \tag{1}$$

This signal when propagated through the vocal tract would result in a speech trace before radiation given by

$$s_l(n) = s_g(n) + s_v(n) \tag{2}$$

$s_g(n)$ and $s_v(n)$ being the contributions of the glottal residual $u_r(n)$ and the vocal tract impulse response $f_v(n)$

$$s_l(n) = u_r(n) * f_v(n) + \delta_g(n) * f_v(n) = s_g(n) + f_v(n) \tag{3}$$

therefore $s_l(n)$ will contain a component $s_g(n)$ contributed by the glottal residual $u_r(n)$ plus the vocal tract response to a train of delta functions $f_v(n)$. Assuming that the vocal tract response and the delayed versions of the glottal residual $u_r(n)$ are fully uncorrelated (second-order decoupling)

$$E\{u_r(n+k) f_v(n)\} = 0; \quad \forall k \tag{4}$$

it could be expected that the glottal component $s_g(n)$ and the vocal component $f_v(n)$ would be equally uncorrelated

$$E\{s_g(n+k) s_v(n)\} = 0; \quad \forall k \tag{5}$$

This property between the vocal and glottal components of radiation-compensated speech would permit the use of Joint-Process Estimation (JPE) to separate one from the other as described in Figure 2.a and Figure 3.

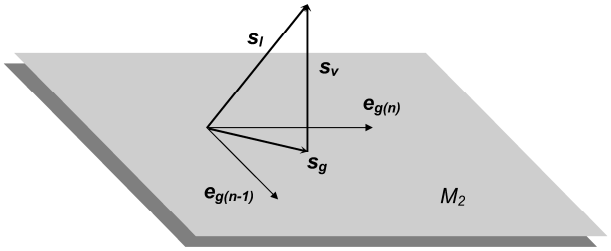

Figure 3. Geometrical interpretation of the separation of the glottal and vocal components. The voiced signal $s_l(n)$ and the devocalized residual error $e_g(n)$ from the vocal tract inversion (see Figure 2.a) are separated under second order statistics by a lattice-ladder filter into the vocal $s_v(n)$ and glottal $s_g(n)$ orthogonal estimates.

In this last figure the results of JPE are interpreted in geometrical terms for a second order process. The delayed versions of the reference signal $e_g(n)$ are used to define an order-P manifold M_P where an estimate of the glottal component s_g is produced by weighting the delayed reference

$$\hat{s}_g(n) = \sum_{i=0}^{P-1} w_i e_g(n-i) \tag{6}$$

which produces an estimation error

$$\varepsilon(n) = s_l(n) - \hat{s}_g(n) \tag{7}$$

By adaptively adjusting the weights w_i an optimal estimate of the vocal component will be obtained by rendering the square absolute value of $\varepsilon(n)$ to a minimum

$$\{w_i\} = \arg\min\{\|\varepsilon(n)\|^2\} \tag{8}$$

therefore obtaining the optimum estimates for the vocal $s_v(n)$ and glottal $s_g(n)$ components at the same time, which could be identified with the orthogonal outputs from the JPE, specifically the results given by (6) and (7) after fulfilling (8).

3. MATERIALS AND METHODS

The separation method described has been checked out using a recording of vowel /a/ by a typical male speaker from a database of 100 normal speakers equally distributed by gender. Subject ages ranged from 19 to 39, with an average of 26.77 years and a standard deviation of 5.75 years. The results shown in Figure 4 correspond to a typical male speaker (label #17F) determined following [7].

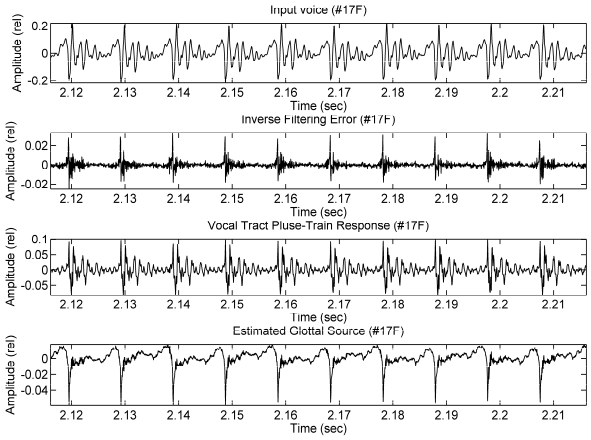

Figure 4. Time domain traces. From top to bottom: input speech $s_l(n)$, inverse filtering error $e_g(n)$, pulse-train response of the vocal tract $s_v(n)$ and glottal estimate $s_g(n)$

The vocal and glottal estimates for the speech trace may be seen in the last two templates, the glottal estimate presenting the expected behaviour as described in the last section. The power spectral densities of each trace are given in Figure 5.

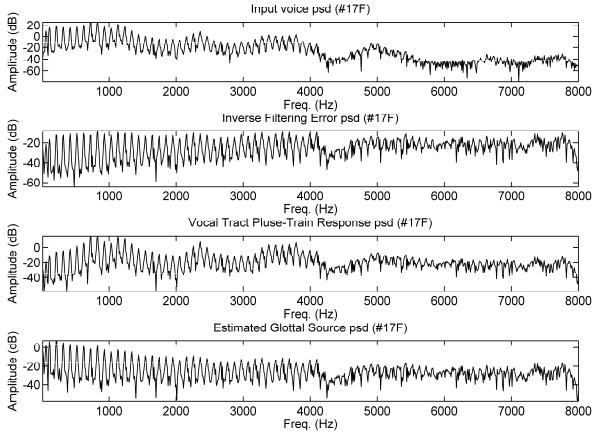

Figure 5. Spectral densities. From top to bottom: input speech $S_l(\omega)$, inverse filtering error $E_k(\omega)$, response of the vocal tract $S_v(\omega)$ and glottal estimate $S_g(\omega)$

In the third template from top it may be appreciated that the formant structure of the vocal tract is equalized with respect to the first template (plain voice), and that the glottal formant has been removed (the glottal formant is the contribution of the glottal source to the power spectral density of voice, which when properly estimated by long-term FFT appears as a hunch-back well below the first vocal tract formant). The harmonic structure and profile of the glottal component is seen in the last template, where the glottal formant can be clearly appreciated around 200 Hz. This characteristic together with some troughs and valleys clearly observable make it differ from the symplistically assumed *1/f* behaviour. The use of the vocal and glottal characteristics in speaker recognition tasks has been established by means of a larger database including a wide representation of the phonetic articulation and the glottal characteristics for 240 speakers taken from [12] which give a good description of intra- and inter-speaker variability. The data set was divided into 176 speakers used for modelling during the training phase, and 64 speakers serving as impostors during the test phase. The training dataset is composed by 10 sentences per each of the modelled speakers comprising approximately 30 sec. of speech. The testing set consisted in 10 sentences per known speaker as well as 10 sentences from each impostor speaker, each sentence lasting 4 sec for both groups. Training and testing sets for each speaker are based on different sentences. The database is phonetically balanced for the Spanish language and although not specifically designed for speaker characterization it has a very rich and complete representation of the acoustic-phonetic variability of each speaker concerning the closed-set speaker identification experiments programmed.

4. RESULTS AND DISCUSSION

The purpose of the present work is two-fold, on one hand to check the proposed methodology to estimate the vocal and glottal components of voiced speech, on the other hand to determine the best parameter templates to improve speaker identification scores. For such purpose eight different parameter templates have been used from the description of the parameterization scheme shown in Figure 2.b. Each speech trace is sampled at 16,000 Hz and subsequently processed to detect the segments showing speech activity, and these are further processed for voicing-unvoicing detection. For voicing fragments the vocal and glottal components $s_v(n)$ and $s_g(n)$ are estimated and integrated. The power spectral density of each trace is estimated by FFT in 512-sample sliding windows. 14 Mel-Frequency Cepstral Coefficients (MFCC) are estimated for the vocal component and 8 MFCC for the glottal component as well. For unvoicing fragments 14 MFCC templates are produced following the same methodology. Pitch and the logarithm of the energy are also computed. Using the available information the following templates were produced for the experiments described:

1. MFCC&P&E: 14 MFCC from raw speech + an estimate of pitch + the logarithm of energy
2. MFCC PG: 8 MFCC from the power spectral density of the glottal source + an estimate of pitch + the logarithm of energy
3. MFCC VT: 14 MFCC from the vocal tract impulse response+ an estimate of pitch + the logarithm of energy
4. MFCC VO: 14 MFCC from the voiced segments + an estimate of pitch + the logarithm of energy
5. MFCC FUSION C: 14 MFCC from raw speech + 14 MFCC from the vocal tract impulse response + 8 MFCC from the glottal source power spectral density + an estimate of pitch + the logarithm of energy

The training session produced Gaussian Mixture Models for each modelled speaker of order $k=\{16, 32, 64, 128, 256, 512\}$ the testing set was processed in a closed-set setup and the scores recorded in relation to the log-likelihood threshold ϑ used for each experiment. The results of the different experiments conducted are given in Figure 6-Figure 10 and in Table 1. The first template (in Figure 6) shows the selection of the baseline DET trace among the results of processing the database parameterized as MFCC&P&E for the GMM orders of $k=\{16, 32, 64, 128, 256, 512\}$. The best choice is determined by minimizing the following functional with respect to k

$$F_k = \int_{\vartheta_1}^{\vartheta_2} f_{AR}(\vartheta,\Gamma_k) f_{RR}(\vartheta,\Gamma_k) d\vartheta \qquad (9)$$

where ϑ is the moving threshold to establish the detection condition, and f_{AR} and f_{RR} are respectively the False Acceptance and False Rejection Rates in terms of the threshold ϑ and the Gaussian Model used Γ_k.

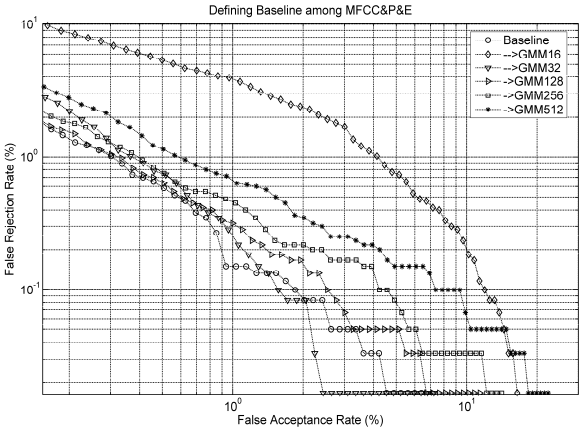

Figure 6. Selection of the best GMM order on comparing the results from parameterization MFCC&P&E.

It may be seen that the selected classifier is the one corresponding to order $k=64$, which will be used as baseline in further comparisons against other parameter settings. In Figure 7, the results of comparing the results from the parameterization of the glottal source against the baseline are presented.

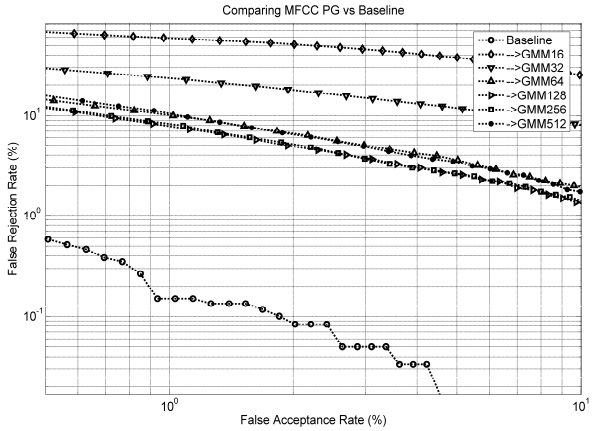

Figure 7. Comparison of the identification results from the Glottal Source parameterization against the baseline.

It may be seen that the glottal parameterization performs rather poorly compared to plain speech, although it shows a certain identification capability. In the next template (Figure 8) the performance of 14 MFCC Vocal Tract + pitch + logE parameters is compared against the Baseline.

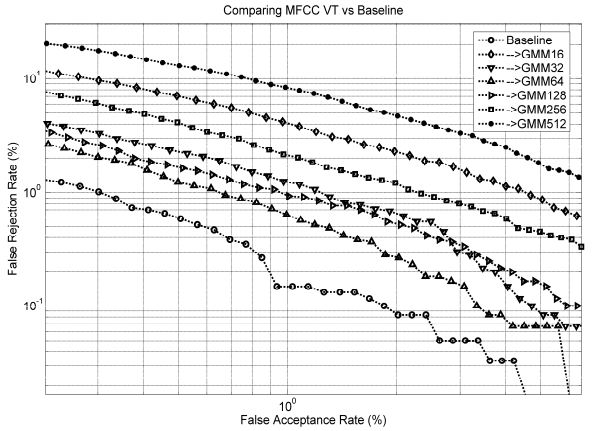

Figure 8. Comparison of the identification results from the parameterization of the vocal tract transfer function against the baseline.

It may be seen that the DET curves from the Vocal Tract Transfer Function show an identification capability *per se*, although their performance is also sensibly worse than that of the Baseline. In the next template (Figure 9) the performance of 14 MFCC for voicing speech (Voicing Only) + pitch + logE is compared against the Baseline.

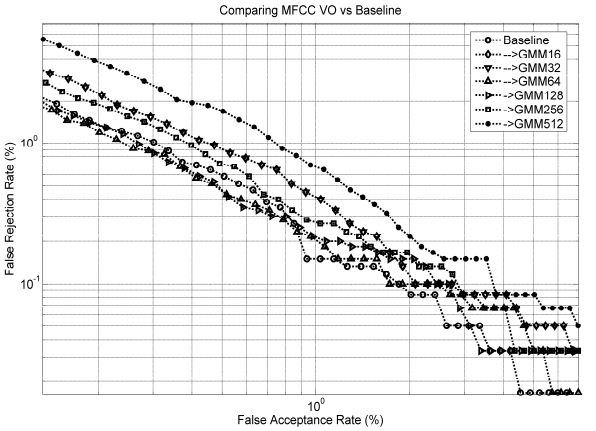

Figure 9. Comparison of the identification results from the parameterization of the voiced segments of speech against the baseline.

In this case it may be seen that some of the detector configurations (*k=64* and *k=128*) may outperform the baseline for certain threshold values, especially those reducing f_{AR}. The last template (Figure 10) shows the combined use of the vocal and glottal estimates mixed with full speech against the Baseline.

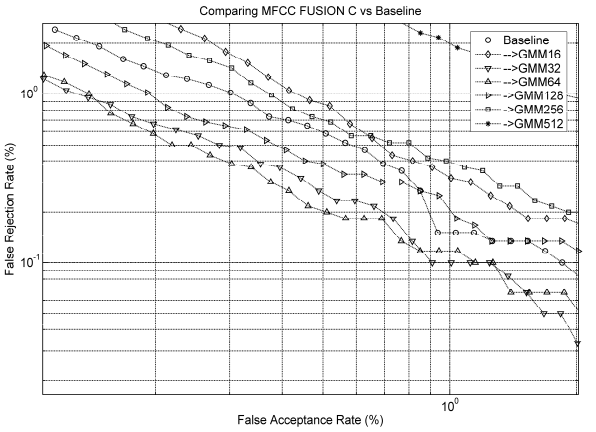

Figure 10. Comparison of the identification results from the combined parameterization of raw speech, vocal tract and glottal source against the baseline.

In this last case a substantial improvement may be observed compared to the Baseline. These same data are summarized in Table 1 which gives the harmonic mean of the point closest to Equal Error Rate conditions for each parameterization template used in the study.

Table 1. Optimal Equal Error Rate (%) conditions for the different parameterization schemes used.

EER/GMM	**16**	**32**	**64**	**128**	**256**	**512**
Baseline	*2.20*	*0.60*	*0.54*	*0.55*	*0.60*	*0.78*
MFCC_VO	*0.68*	*0.68*	*0.48*	*0.50*	*0.59*	*0.84*
MFCC_VT	*2.13*	*1.13*	*0.81*	*0.97*	*1.49*	*3.20*
MFCC_GS	*15.94*	*8.82*	*4.05*	*3.41*	*3.35*	*3.90*
MFCC_FS	*0.62*	*0.36*	*0.35*	*0.42*	*0.57*	*1.40*

It may be seen that the best baseline scores are *0.54%*, the glottal estimate alone behaving far worse (*3.41%*). The vocal estimate alone renders better results than the glottal estimate (*0.81%*) possibly due to the elimination of the glottal features from acoustic-phonetic information (showing large intra-speaker variability), which allows a better estimation of message-dependent features from spectral characteristics. Nevertheless if glottal and vocal characteristics are estimated independently and added as an extension to raw speech estimates (MFCC FUSION C), the overall scores are substantially better, reducing the optimum EER to 65% the corresponding baseline result. It is especially interesting to observe the rather low False Rejection rate produced under 1% False Acceptance (of around 0.125% for GMM orders 32 and 64 accordingly to Figure 10), which would make this technique especially suitable for secure access applications.

5. CONCLUSIONS

At this point it should be worth to comment the use of MFCC's both for the characterization of the vocal tract and glottal source spectral profiles. Cepstral parameterization has been justified by its separation capability of the source and filter parts of voice. This being true, there is another reason to justify the use of this parameterization, which is its implicit robustness and its successful use in voice characterization for pathology studies [13]. Its use in the present case is supported by these considerations as well as by the accuracy of the results obtained. Accordingly with what has been exposed, using the speech generation model to derive differentiate speech features apparently seems to work well under the limitations of the experiments and the database used. The decomposition of voiced speech into vocal and glottal source estimates seems to produce more accurate and independent templates helping to improve the False Rejection rates substantially. Although the reconstructed glottal estimate is not optimum under the Liljencrants-Fant (LF) [14] model and should not be used in glottal studies, it seems that the biometric information extracted from the glottal estimate can help substantially in improving the behaviour of speaker identification systems confirming the conclusions in the work of Plumpe et al [6]. The main difference of the present approach to Plumpe's resides in the decomposition methodology used granting the orthogonality of the templates combined, and in basing the parameterization in the frequency domain rather than in the time domain. The decomposition methodology based on Joint-Process Estimation seems to work well using a criterion derived from second-order statistics, although the estimation of the reference signal may be quite important for the application of this method and some more strategies should be further investigated. Other possible approaches based on higher-order statistics as Independent Component Analysis [15] could also be used, this matter pending of further research.

ACKNOWLEDGMENTS

This work is being funded by grants TIC2003-08756 and TEC2006-12887-C02-01/02 from Plan Nacional de I+D+i, Ministry of Education and Science, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http.//www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

REFERENCES

[1] Alku, P., "Parameterisation Methods of the Glottal Flow Estimated by Inverse Filtering", *Proc. of VOQUAL'03*, Geneva, August 27-29, 2003, pp. 81-87.

[2] Holmberg, E. B., Hillman, R. E., and Perkell, J. S., "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", *J. Acoust. Soc. Am.*, Vol. 84, No.2, 1988, pp. 511-529.

[3] Price, P.J., "Male and female voice source characteristics: Inverse filtering results", *Speech Communication*, Vol. 8, 1989, 261-277.

[4] Sulter, A. R., and Wit, H. P., "Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age", *J. Acoust. Soc. Am.,* Vol. 100, 1996, pp. 3360-3373.

[5] Kuo, J., Holmberg, E. B., Hillman, R. E., "Discriminating Speakers with Vocal Nodules Using Aerodynamic and Acoustic Features", *Proc. of the ICASSP'99*, Vol. 1, 15-19 March 1999, pp. 77-80.

[6] Plumpe, M. D., Quatieri, T. F., Reynolds, D. A., "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification", *IEEE Trans. on Speech and Audio Proc.*, Vol. 7, No. 5, 1999, pp. 569-586.

[7] Gómez, P., et al., "Biometrical Speaker Description from Vocal Cord Parameterization", *Proc. of ICASSP'06*, Toulouse, France, 2006, pp. 1036-1039.

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

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

[10] 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.

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

[12] A. Moreno et al., "ALBAYZIN Speech Database: Design of the Phonetic Corpus," *Proc. Eurospeech'93*, vol. 1, 1993, pp. 175–178.

[13] 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.

[14] Fant G., Liljentcrants J., Lin Q., "A four-parameter model of glottal flow", *STL-QSPR*, Vol. 4, 1985, pp 1-13.

[15] Gómez, P. et al., "DOA Detection from HOS by FOD Beamforming and Joint-Process Estimation", *Proc. of the Fifth Int. Conf. on Independent Component Analysis – ICA'04*, LNCS, Vol. 3195, 2004, pp. 824-831.