Gender Detection in Running Speech from Glottal and Vocal Tract Correlates

Cristina Muñoz-Mulas, Rafael Martínez-Olalla, Pedro Gómez-Vilda,
Agustín Álvarez-Marquina, and Luis Miguel Mazaira-Fernández

Neuromorphic Speech Processing Lab, Centro de Tecnología Biomédica,
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n,
28223 Pozuelo de Alarcón, Madrid
ce.munoz@upm.es

Abstract. Gender detection from running speech is a very important objective to improve efficiency in tasks as speech or speaker recognition, among others. Traditionally gender detection has been focused on fundamental frequency (f0) and cepstral features derived from voiced segments of speech. The methodology presented here discards f0 as a valid feature because its estimation is complicate, or even impossible in unvoiced fragments, and its relevance in emotional speech or in strongly prosodic speech is not reliable. The approach followed consists in obtaining uncorrelated glottal and vocal tract components which are parameterized as mel-frequency coefficients. K-fold and cross-validation using QDA and GMM classifiers showed detection rates as large as 99.77 in a gender-balanced database of running speech from 340 speakers.

Keywords: speech processing, joint-process estimation, speaker's biometry, contextual speech information.

1 Introduction

Accurate gender detection from voice is a very important premise in many speech and voice analysis tasks, as automatic speech recognition (ASR), voice pathology detection (VPD), automatic speaker characterization (ASC) or speech synthesis (SS). It is well known that many applications improve substantially detection error trade-offs or classification and recognition rates if appropriate gender-oriented models are used, as inter-speaker variability is reduced. This is especially so in voice quality analysis for organic pathology detection [1]. For such pitch estimates were classically used as it was thought that pitch is a precise mark of gender, when actually it is not. It is true that pitch in modal phonation (that one produced under quiet and controlled conditions in sustained vowels as /a/ as more comfortably as possible) tends to distribute differently in male and female voices. But these conditions are not fulfilled in running speech, where pitch may be altered by prosody and emotion effects, or in singing. Voice pathology may alter also pitch, reducing the fundamental frequency (f0) in females or incrementing it in males, and phonation bifurcations may produce

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 25–32, 2013.

drastic changes in pitch within an octave. Other factors maybe the interaction between the glottal formant and the first vocal tract formant, and the influence of telephone channels in affecting the fundamental frequency band. Therefore detecting gender based on a single feature as f0 may become rather unreliable having in mind the problems associated to f0 estimation in itself, especially if a wider description of biometric features as gender and age is involved. There are several gender detection techniques which are of interest to this study. A classical analysis is given in [2, 3], in which the authors investigate the relative role of fundamental frequency and formants in gender recognition experiments using mainly vowels. The results claimed 100% accuracy with a limited number of speakers (27 male and 25 female). In [4] gender detection is based on a combination of features derived only from the glottal source separated from vowel segments by inverse filtering and approximate reconstruction. The features used are f0, the instant of maximum glottal area (gap), the maximum derivative of the gap, the slope of the glottal flow spectrum, and the harmonic amplitude ratios. False classification rates are 5.3% for males and 4.1% for females on a database with 92 speakers (52 male and 40 female). Several inconveniences are found in all these approaches. The first one is to rely strongly on f0 estimates, having in mind that this is a complicate task, or to depend on estimates of the formants, which are also dependent on f0 and on peak tracking. These facts raise the question of providing gender detection based on the following premises: exclude f0 as feature if possible; step on robust features derived from vocal tract and glottal source estimates obtained from a source-filter separation technique granting orthogonal descriptions; use running speech similar to what can be found in real applications; test the methodology on a large database enough to grant statistical significance. The present approach is based on a careful reconstruction of the glottal source and resonant cavities using techniques derived from voice pathology studies [5]. The paper is organized as follows: in Section 2 a description of the methodology to produce statistically independent features for the vocal tract and the glottal source. In Section 3 the database used in the experiments is described and the experimental setup is commented. Section 4 is devoted to present and discuss gender detection results obtained using the methodology and database described. In section 5 conclusions are derived.

2 Present Approach

The model of speech production proposed by Fant is a very well know one to need any further explanation [6] (see Fig. 1). Its main interest is founded in the presence of an excitation which may be voiced or voiceless, modified by a time-varying filter representing the articulation organs (pharynx, oral and nasal cavities), usually modeled as a tube of irregular shape which may be approximated by a concatenation of time-varying cross-section tubes. In a first order approach the system is considered loss-less, and time variations are handled by means of adaptive algorithms which may cope with changes in the cross-section profile.

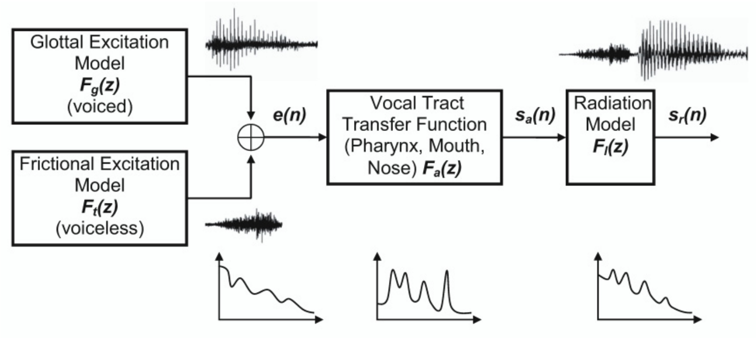

Fig. 1. Fant's source-filter model of speech production: the excitation signal e(n) may be produced by phonation (voicing) or by turbulent excitation (voiceless). The articulation organs (pharynx, vocal/nasal tracts) have a specific behavior in the frequency domain given as a set of resonances and anti-resonances (mid-bottom) which produce a pre-radiated speech signal $s_a(n)$. The radiation model changes the spectral tilt of produced speech $s_r(n)$.

The interest of the model resides in the possibility of obtaining features to describe separately the glottal source (in voiced sounds) and the vocal tract filter (both in voiced and in voiceless sounds), thus a descriptor of the human features behind the vocal tract will be available in any situation where speech is present. The glottal source in voiced sounds is affected by the length, mass and tension of the vocal folds, which are clearly differentiated by gender (longer length, higher mass and lower tension in adult males with respect to females). The vocal tract is also clearly differentiated accordingly with gender (overall length and pharyngeal cavity dimensions [7]), thus a second set of features may be added to those from the glottal source for detection purposes. Traditionally the separation of the source and filter have been carried out by inverse filtering using estimates of the vocal tract structure to remove the resonances introduced by its equivalent transfer function in speech spectra. This separation has taken into account source-system coupling effects mainly. In the present approach a joint-process estimation methodology is proposed to create orthogonal estimates of the glottal source and vocal tract impulse responses under second order statistics [8]. The combined joint-process estimator consists in a lattice adaptive filter and a ladder mirror filter, both using dynamic adaptation of weights to produce residual signals which may be shown to be uncorrelated under second order statistics (see Fig. 2). The source-filter separation method (a) consists in producing a first estimate of the inverse vocal tract transfer function $H_v(z)$, which is used to estimate a de-vocalized residual error $e_g(n)$. Classically this residual was considered useless [9] to be recently recognized as an important source of information on phonation characteristics [5, 9]. This residual is contrasted in a lattice-ladder joint-process estimator against the radiation-compensated speech $s_l(n)$ to produce two other estimates $s_g(n)$ and $s_v(n)$, corresponding to the glottal and tract components. These correlates present the property of being orthogonal under second-order statistics [8], and are used in (b) to produce mel-frequency estimates of the vocal and glottal power spectral densities. These vectors are aligned with estimates of pitch (f0), voiced/voiceless (v/u) and the log of the energy to define the final feature vector. One of the most relevant aspects of parameterization is to decide on the orders of mfcc sets of parameters for the vocal tract (k_v) and glottal source (k_g) in Fig. 2.

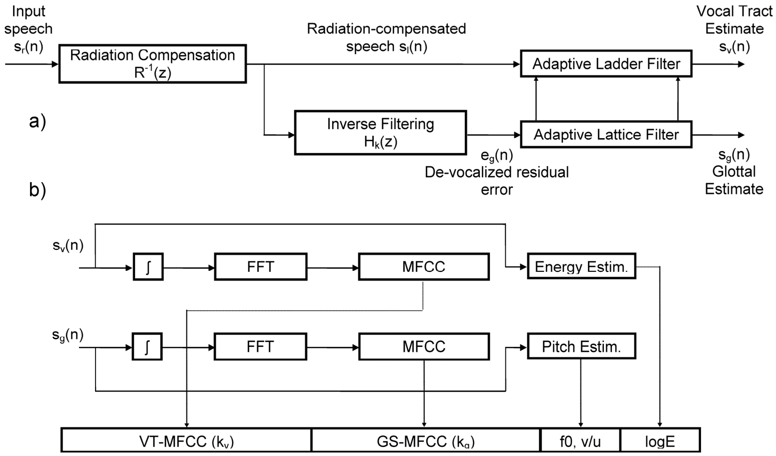
Input speech $s_r(n)$
Radiation Compensation $R^{-1}(z)$
Radiation-compensated speech $s_l(n)$
Vocal Tract Estimate $s_v(n)$
Adaptive Ladder Filter
a)
Inverse Filtering $H_k(z)$
Adaptive Lattice Filter
$e_g(n)$ De-vocalized residual error
$s_g(n)$ Glottal Estimate
b)
$s_v(n)$
∫
FFT
MFCC
Energy Estim.
$s_g(n)$
∫
FFT
MFCC
Pitch Estim.
VT-MFCC (k_v)
GS-MFCC (k_g)
f0, v/u
logE

Fig. 2. Parameterization method: a) Lattice-Ladder Adaptive Joint-Process estimator to separate source and filter estimates $s_g(n)$ and $s_V(n)$; b) mel-frequency cepstral parameterization of glottal and tract components

There is not a clear criterion on this respect except considering the number of frequency channels to split spectra following mel scale. In general the region of interest of the vocal tract transfer function extends well to 8 kHz, whereas the relevant glottal information concentrates mainly in the band 0-5 kHz. Therefore a 20-band mfcc parameterization was used both to parameterize full speech and the vocal tract transfer function (k_V=20), whereas a 12-band mfcc parameterization was used for the glottal source (k_g=12), at a sampling rate of 16 kHz. This strategy creates a 55-dimmension feature vector. Examples of feature distributions from the database used, which will be described in section 3 are given in Fig. 3. Obviously not all the features will have the same relevance accordingly to gender detection criteria, therefore a study of parameter relevance would be mandatory. This is carried out using Fisher's metric according to:

$$F_m = \frac{\left(\bar{\xi} - \bar{\xi}_f\right)^2 n_f + \left(\bar{\xi} - \bar{\xi}_m\right)^2 n_m}{\operatorname{var}\left(\xi_f\right)\left(n_f - 1\right) + \operatorname{var}\left(\xi_m\right)\left(n_m - 1\right)} \tag{1}$$

where ξ is the feature vector for the whole speaker's set, $\bar{\xi}_m$ and $\bar{\xi}_f$ are the respective average feature vectors for male and female speaker sets, and n_m and n_f are the respective number of speakers in each set. The list including the most relevant parameters in the feature set is given in Table 1. The analysis exposed in the table is very clarifying concerning feature selection: among the 14 first features by Fisher's metric the two most relevant ones are glottal source related (10 and 8); f0 is classified in third place, the following six ones are also glottal source related (11, 12, 7, 9, 1 and 6), the most relevant one derived from full speech is in position 10, and its Fisher's metric is almost four times lower than the first one.

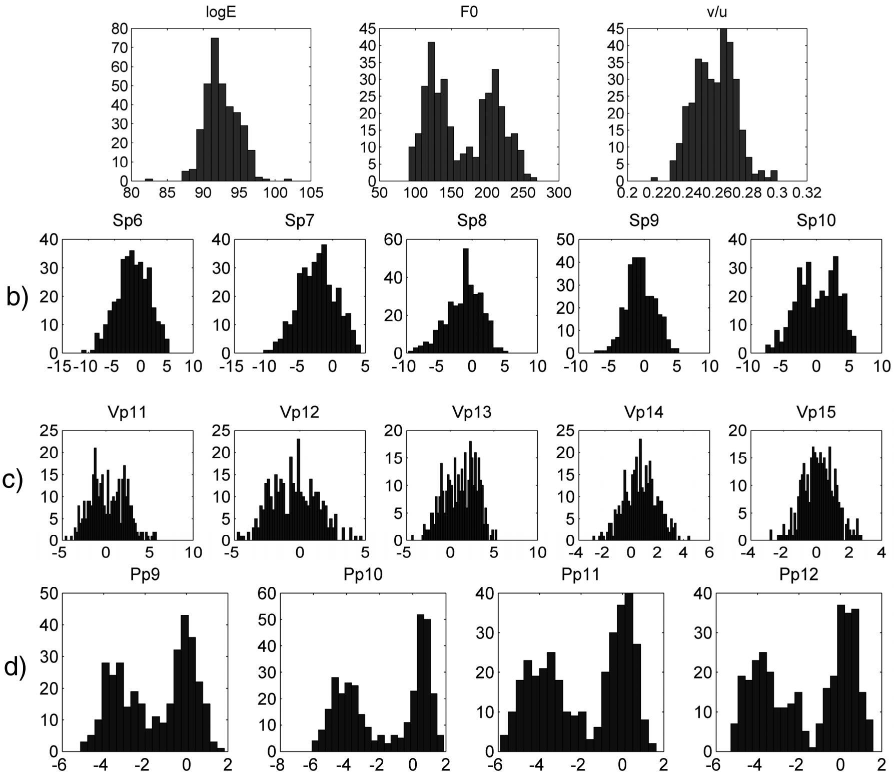

Fig. 3. Distribution histograms for some of the features estimated: a) energy, f0 and voiced/unvoiced; b) selected speech mfcc's (Sp); c) selected vocal tract mfcc's (Vp); d) selected glottal source mfcc's (Pp). Some distributions show clear gender bimodality (f0, Sp10, Vp11, Pp9, Pp10, Pp11, Pp12).

Table 1. Fisher's metric for a subset of estimated features

Order	Feature	Value	Order	Feature	Value
1	Pp10	2.0201	9	Pp6	0.7938
2	Pp8	1.6473	10	Sp10	0.5540
3	f0	1.4616	11	Vp13	0.5496
4	Pp11	1.4455	12	Vp11	0.5281
5	Pp12	1.2388	13	Pp5	0.4603
6	Pp7	1.2178	14	Sp13	0.3867
7	Pp9	0.9271	32	v/u	0.0563
8	Pp1	0.8136	49	logE	0.0019

The first feature from vocal tract is in position 11 (Vp13). Finally the feature voiced/unvoiced (v/u) and logE have been included as a reference in positions 32 and 49. These results do not clarify possible redundant relations among the different features, therefore in detection tasks instead of the original feature vector its PCA transformation has been used in the experiments.

3 Materials and Methods

The database used is a classical benchmark for running speech in Spanish [10]. It is composed of recordings from 340 speakers balanced by gender (170 males and 170 females), distributed by age in the range from 18 to 64 years. Half of the speakers were under 30. Each speaker was entitled to produce at least 25 sentences lasting from 2 to 4 s long comprising the complete phonetic repertoire of central peninsular standard dialect, supposedly balanced in contents and co-articulation. The database was recorded with high quality standards in 16 kHz and 16 bits (suppression of low frequency noise under 16 Hz, HQ microphones, equalization, direct digital recording, sound proof room). It comprises three corpora: phonetic, geographic and Lombard. The amount of speech from the phonetic corpus used in the experiments described is over 30,000 s. Two classifiers have been compared in separating speakers by gender, the first one is based on quadratic discriminant analysis (QDA). The second classifier is a classical Gaussian mixture model (GMM) of order 2 (one per gender) [10]. Both classifiers operated on the PCA transformed feature vector. Two types of tests were designed. In one of them the experiments carried out with both classifiers were organized as 5 random cross-validation tests in which the database was divided by speakers (equally balanced) in two subsets including 40% of the speakers for training and 60% for testing. Random speaker selection was used to fit the train and test sets in each experiment. In the second type of experiment the speakers set was divided in 5 subsets comprising 20% of the speakers, equally balanced by gender. Each experiment used one of the subsets for training and the four remnant sets for testing, therefore each experiment was configured with 20% of the dataset for training and 80% of the dataset for testing. Detection was speaker-based.

4 Results and Discussion

Blind clustering by MANOVA analysis was carried out to determine the relevance of each feature set. The results are given in Fig. 4 in terms of the first two canonical components from MANOVA (c_2vsc_1): (a) if speech mfcc's are used two main clusters are observed which are clearly separated with an overlapping region within dot lines (6 errors); (b) vocal tract mfcc’s reduce the errors (4) but clusters are less separate; (c) glottal source mfcc’s separate clusters better, but number of errors is larger (8); (d) if glottal pulse and vocal tract mfcc's are combined the separation between clusters improve and the number of errors is lower (3). In the detection experiments using QDA and GMM’s reported in Table 2 the feature vector was configured in seven different ways: 20 speech-derived features only (S), 12 glottal source-derived features (GS), 20 vocal tract-derived features (VT), 32 speech and glottal source fused features (S+GS), 40 speech and vocal tract fused features (S+VT), 32 glottal source and vocal tract features (GS+VT) and 52 speech, glottal source and vocal tract fused features (S+GS+VT). The detection rates obtained in each task after averaging over the 5 experiments per trial are given in Table 2. These results confirm the first impression derived from MANOVA analysis: complementing speech features with glottal and vocal tract features outperform features from original speech only.

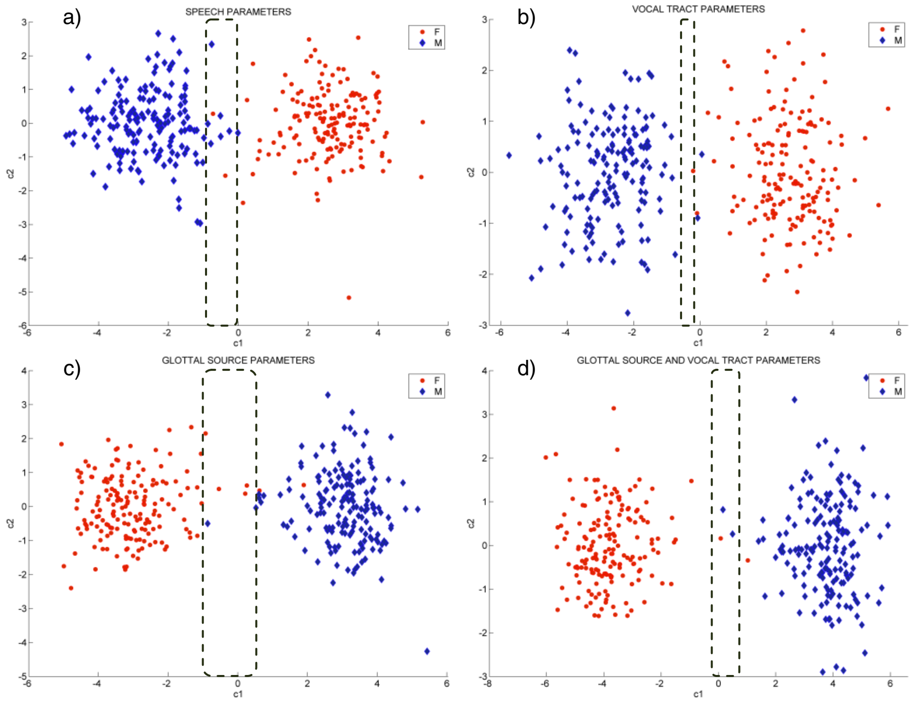

Fig. 4. Gender group separation after MANOVA analysis on different mfcc combinations from: a) original speech, 6 OS); b) vocal tract transfer function (TF), 4 OS; c) glottal source power spectral density (PSD), 8 OS; d) glottal source PSD and vocal tract TF; 3 OS. OS: overlapped subjects.

Table 2. Detection results: averages over 5 experiments

Av. Rel. Err.	S	GS	VT	S+GS	S+VT	GS+VT	S+GS+VT
QDA-xval	98.47	98.00	98.24	99.18	98.41	98.94	99.59
QDA-5 fold	96.08	96.08	96.08	98.53	98.53	98.53	99.41
GMM-xval	97.18	98.65	99.24	99.36	98.06	99.29	**99.77**
GMM-5 fold	99.18	99.41	99.35	99.41	99.47	98.94	99.59

5 Conclusions

First of all it must be stressed that f0, although estimated and put in comparison vs other features in Table 1 is not used in gender detection, as it may be inferred from the feature sets used in the experiments. The intention in proceeding so was two-fold, on one hand to avoid the problems found in accurate pitch estimation, on the other hand to avoid intra-speaker dispersion due to prosody and emotional factors (especially in male). Accordingly to the results this decision has shown to be crucial in obtaining reliable and robust results. From what has been exposed the following conclusions may be derived:

- The estimation of de-correlated components of the vocal tract and glottal source seems to be well supported theoretically and by experimentation.
- Mel-frequency cepstral features of the vocal tract impulse response and glottal source spectral densities can be considered robust descriptors of phonation and articulation gestures for both genders in running speech.
- GMM classifiers performed better than QDA's, especially using cross-validation, although 5-fold validation results were over 99% for all feature combinations.
- Vocal tract features did not perform as well as the ones from glottal source.
- Glottal source features outperformed speech-derived ones. A possible explanation for this behavior could rely on lesser dependence of articulation.
- Combinations of glottal source with speech derived features outperformed other combinations except the combination of the three kinds of parameters. This fact needs further investigation.

Future lines are to extend this methodology to the classification of speakers by age, considering that the glottal source is very much influenced by aging as well, and to non-modal speech corpora (emotional speech, singing, pathological speech).

Acknowledgments. Work funded by grants TEC2009-14123-C04-03 and TEC2012-38630-C04-04, Ministry of Econ. Affairs and Compet., Spain.

References

1. Fraile, R., Saenz-Lechon, N., Godino-Llorente, J.I., Osma-Ruiz, V., Fredouille, C.: Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex. Folia Phoniatrica et Logopaedica 61, 146–152 (2009)
2. Wu, K., Childers, D.G.: Gender recognition from speech. Part I: Coarse analysis. J. Acoust. Soc. Am. 90(4), 1828–1840 (1990)
3. Childers, D.G., Wu, K.: Gender recognition from speech. Part II: Fine analysis. J. Acoust. Soc. Am. 90(4), 1841–1856 (1991)
4. Sorokin, V.N., Makarov, I.S.: Gender recognition from vocal source. Acoust. Phys. 54(4), 571–578 (2009)
5. Gómez, P., Fernández, R., Rodellar, V., Nieto, V., Álvarez, A., Mazaira, L.M., Martínez, R., Godino, J.I.: Glottal Source Biometrical Signature for Voice Pathology Detection. Speech Comm. 51, 759–781 (2009)
6. Fant, G.: Acoustic theory of speech production. Walter de Gruyter (1970)
7. Titze, I.: Principles of voice production. Prentice Hall, Englewood Cliffs (1994)
8. Manolakis, D., Ingle, V.K., Kogon, S.M.: Statistical and Adaptive Signal Processing. Artech House (2005)
9. Prasanna, S.R.M., Gudpa, C.S., Yegnanarayana, B.: Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Communication 48, 1243–1261 (2006)
10. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Nadeu, C.: Albayzin Speech Database: Design of the Phonetic Corpus. In: Proc. Eurospeech 1993, vol. 1, pp. 653–656 (1993)
11. Reynolds, D., Rose, R.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. SAP 3(1), 72–83 (1995)