ON THE MUTUAL INFORMATION OF GLOTTAL SOURCE ESTIMATION TECHNIQUES FOR THE AUTOMATIC DETECTION OF SPEECH PATHOLOGIES

T. Dubuisson, T. Drugman, T. Dutoit
TCTS Lab, Faculté Polytechnique de Mons, 31 Boulevard Dolez, 7000 Mons, Belgium

Abstract: **This paper focuses on the automatic detection of speech pathologies by exploiting the estimation of the glottal source. Three methods of estimation are compared and time and spectral features are extracted. The relevancy of these features is assessed by means of information theory-based measures. This allows an intuitive interpretation in terms of discrimination power and redundancy between the features. It is discussed which features are informative or complementary for detecting voice pathologies and the glottal source estimation methods are compared.**

Keywords: **Voice Pathology, Glottal Source, Mutual Information**

I. INTRODUCTION

Perceptive evaluation performed by clinicians suffers from the dependency on the experience of the listener and the inter- and intra-judges variability. There is thus a need to develop objective tools. For this, a part of research in speech processing has focused on the detection of speech pathologies from audio recordings. Indeed it could be useful to detect disorders when perturbations are still weak, to prevent the degradation of the pathology, or to measure the voice quality before and after surgery [1]. As video recordings of the vocal folds show that their behavior is linked to the perception of different kinds of voice qualities, including pathologies, isolating and parametrizing the glottal excitation should lead to a better discrimination between normal and pathological voices. Such parametrizations of the glottal pulse have already been proposed both in time and frequency domains ([2], [3]).

This paper pursues the work presented in [4], in which it was shown that features respectively extracted from the vocal tract and glottal contributions (estimated by the IAIF algorithm [5]) are synergic and can lead together to an efficient discrimination of voice disorders. The present study addresses the comparison between IAIF and two other methods for the same problem. As in [4], the performance of classification is assessed by computing information theory-based measures in order to provide an intuitive interpretation in terms of discrimination power, redundancy and synergy between the features.

The paper is structured as follows. In Section 2, the different methods of glottal source estimation are presented. Section 3 defines the features extracted from the glottal source. Section 4 reviews the mutual information-based measures that are used in this work and highlights their interpretation for a classification problem. Experiments and results are detailed in Section 5. It is discussed which features are informative for the detection of voice disorders and which ones are complementary. Finally Section 6 concludes.

II. GLOTTAL SOURCE ESTIMATION

Three methods of glottal source estimation are considered here: the Complex Cepstrum Decomposition (CCD) [6], the Iterative Adaptative Inverse Filtering (IAIF) [5] and the Closed Phase Inverse Filtering (CPIF) technique [7]. The application of these three methods on a fragment of a normophonic sustained vowel /a/ is presented in Fig. 1.

A. Complex Cepstrum Decomposition

It has been recently shown that complex cepstrum can be efficiently used for glottal source estimation [6].This method aims at separating the minimum and maximum-phase components of the speech signal. Indeed it has been shown previously [8] that speech is a mixed-phase signal where the maximum-phase (i.e. anti-causal) contribution corresponds to the glottal open phase, while the minimum-phase component is related to the vocal tract transmittance (assuming an abrupt glottal return phase). Isolating the maximum-phase component of speech then provides a reliable estimation of the glottal source, which can be achieved by the complex cepstrum.

B. Iterative Adaptative Inverse Filtering

The IAIF technique [5] (publicly available in the Aparat Toolkit [9]) iteratively estimates the vocal tract contribution from the speech signal using a Discrete All Pole model whose order is different for the successive iterations. The glottal source is estimated by filtering the speech signal by the inverse of the filter modeling the contribution of the vocal tract.

C. Closed Phase Inverse Filtering

The CPIF technique exploits the fact that the glottal cycle consists of two phases, during which the vocal folds are respectively open and closed [7]. The key idea of this technique is to estimate the vocal tract transmittance during the closed phase, when it is assumed to be almost free of any excitation. Linear prediction is thus applied on

the speech signal during the closed phase and the glottal source is estimated by inverse filtering of the speech signal.

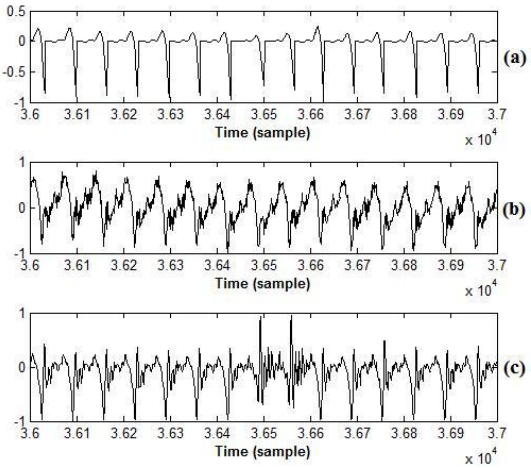

Fig. 1. Comparison of the three glottal source estimations (a: CCD; b: IAIF; c: CPIF) for a normophonic sustained vowel.

III. Feature extraction

Features are here extracted from glottal pitch-synchronous frames in voiced parts of speech. These frames are two-pitch period long, centered on the glottal closure instant (GCI) and weighted by a Blackman window. Pitch and voicing decision are computed using the Snack library [10] while GCIs are located according to the method described in [11].

A. Spectral Features

The amplitude spectrum of a voiced glottal source generally presents a low-frequency response called *glottal formant* produced during the open phase [3]. This formant is here characterized both by its frequency F_g and bandwidth B_w.

The spectral content of the glottal source spectrum is summarized by computing characteristics describing the repartition of its energy. The global repartition of spectral energy is captured in the spectral center of gravity *CoG*. A finer distribution of energy is quantified by considering an approach similar to [12] but using the perceptive mel scale. For this, the power spectral density is weighted by a mel-filterbank consisting on 24 triangular filters equally spaced along the whole mel scale. Three perceptive spectral balances are then computed:

$$Bal_1 = \frac{\sum_{i=1}^{4} PE(i)}{\sum_{i=1}^{24} PE(i)} \; Bal_2 = \frac{\sum_{i=5}^{12} PE(i)}{\sum_{i=1}^{24} PE(i)} \; Bal_3 = \frac{\sum_{i=13}^{24} PE(i)}{\sum_{i=1}^{24} PE(i)} \quad (1)$$

where PE(i) denotes the cumulated weighted power spectral density for the ith filter.

B. Time Features

In many studies, the glottal flow and its derivative (called here *glottal source*) have been used to characterize voice quality [2]. Two parameters are computed here for characterizing the amplitude and duration of the open phase of the glottal cycle. The Normalized Amplitude Quotient (*NAQ*) [2] is defined as the ratio of glottal flow amplitude and the minimum peak of glottal flow derivative, normalized by the length of the glottal cycle. The Quasi-Open Quotient (*QOQ*) [2] is defined as the duration during which the glottal flow is 50% above the minimum flow. Unlike the other parameters, *NAQ* and *QOQ* are computed for one glottal cycle instead of a two-period long frame centered on the GCI. Furthermore, we observed in [4] that the discontinuity at the GCI is generally more significant in case of normal voice than in case of pathological voice. The minimum value at the GCI (*minGCI*) of energy-normalized glottal source frames is thus also considered here.

IV. Information theory-based measures

The problem of automatic classification consists in finding a set of features X_i such that the uncertainty on the determination of classes C is reduced as much as possible [13]. For this, Information Theory [14] allows to assess the relevance of features for a given classification problem, by making use of the following measures (where $p(.)$ denotes a probability density function):

- The entropy of classes C is expressed as:

$$H(C) = -\sum_c p(c) \log_2 p(c) \quad (2)$$

and can be interpreted as the amount of uncertainty on their determination.

- The mutual information between one feature X_i and classes C:

$$I(X_i; C) = \sum_{x_i} \sum_c p(x_i, c) \log_2 \frac{p(x_i,c)}{p(x_i)p(c)} \quad (3)$$

can be viewed as the information the feature X_i conveys about the considered classification problem, i.e. the discrimination power of one individual feature.

- The joint mutual information between two features X_i, X_j, and classes C can be expressed as:

$$I(X_i, X_j; C) = I(X_i; C) + I(X_j; C) - I(X_i; X_j; C) \quad (4)$$

and corresponds to the information that features X_i and X_j, when *used together*, bring to the classification problem. The last term can be written as:

$$I(X_i; X_j; C) = \sum_{x_i}\sum_{x_j}\sum_{c} p(x_i, x_j, c)\log_2 \frac{p(x_i,x_j)p(x_i,c)p(x_j,c)}{p(x_i,x_j,c)p(x_i)p(x_j)p(c)} \quad (5)$$

An important remark has to be underlined about the sign of this term. It can be noticed from expression of $I(X_i, X_j; C)$ that a positive value of $I(X_i; X_j; C)$ implies some **redundancy** between the features, while a negative value means that features present some **synergy** (depending on whether their association brings respectively less or more than the addition of their own individual information).

V. Experiments

A. Database

A popular database in the domain of speech pathologies is the MEEI Disordered Voice Database [15]. This database contains sustained vowels and reading text samples, from 53 subjects with normal voice and 657 subjects with a large panel of pathologies. Here, all the sustained vowels of the MEEI Database resampled at 16 kHz are considered.

B. Mutual Information Computation

To evaluate the significance of the proposed features, the following measures are computed:

- the relative intrinsic information of one individual feature $I(X_i; C)/H(C)$, i.e. the percentage of relevant information conveyed by the feature X_i,
- the relative redundancy between two features $I(X_i; X_j; C)/H(C)$, i.e. the percentage of their common relevant information,
- the relative joint information of two features $I(X_i, X_j; C)/H(C)$, i.e. the percentage of relevant information they convey together.

For this, equations presented in Section IV are calculated. Probability density functions are estimated by a histogram approach. The number of bins is set to 50 for each feature dimension, which results in a trade-off between an adequately high number for an accurate estimation, while keeping sufficient samples per bin. Since features are extracted at the frame level, a total of 32000 and 107000 examples is available respectively for normal and pathological voices. Mutual information-based measures can then be considered as being accurately estimated. Class labels correspond to the presence or not of a voice disorder.

C. Results

The values of the measures detailed in the previous section for the three methods are presented in Fig. 2. For each table, the diagonal indicates the percentage of relevant information conveyed by each feature. It can be observed that *QOQ* is the most informative feature for CPIF and CCD methods (respectively 31.5% and 32.8%) while F_g is slightly more informative (25.9%) than *QOQ* in the case of IAIF method. The top-right part contains the values of relative joint information of two features. When used together, the combination of *QOQ* and F_g brings, for the three methods, the most important information about the classification problem, with a maximum value for the CPIF method (63.8%). The bottom-left part shows the values of relative redundancy between two features. For CCD and CPIF methods, F_g is synergic ($I(X_i; X_j: C) < 0$) with all the features, including QOQ, while this latter is less synergic and in some cases redundant with the other features.

The results show that applying the CCD technique gives generally better results than other methods in terms of intrinsic discrimination power. The synergy for the CDD technique is also the highest for most of features pairs. Moreover, using the combination of *QOQ* and F_g computed by CCD is the most interesting for the distinction between normal and pathological voices. Indeed, their mutual information is high, each feature brings its own information in the combination and is not redundant with the information conveyed by the other.

For the three methods, the highest amount of information conveyed by the combination of two features is about 60%. This means that there is a need of other information to distinguish normal and pathological voices. For this, it was shown in [4] that combining only one vocal tract-based and one glottal feature allows explaining 81% of the difference between normal and pathological voices.

VI. Conclusion

This paper focused on the problem of automatic detection of voice pathologies from the speech signal. The goal was to compare the classification performance of the features extracted from the glottal source estimated by three different methods (CCD, IAIF, and CPIF). These features were assessed through mutual information-based measures. It turned out that CCD technique generally provides features that convey higher intrinsic, mutual information and synergy. It was also shown that the couple of features (*QOQ*, F_g) has the highest mutual information (63.8%) and is also characterized by a high synergy, meaning that their association brings more than the addition of their intrinsic information.

ACKNOWLEDGEMENTS

The authors thank the Walloon Region, Belgium, for its support (grant WALEO II ECLIPSE #516009). This paper presents research results of the Belgian Network DYSCO, funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors. Thomas Drugman is supported by the "Fonds National de la Recherche Scientifique" (FNRS).

CCD	Fg	Bw	CoG	Bal1	Bal2	Bal3	MinGCI	Naq	Qoq
Fg	**12,0**	42,0	47,8	35,7	41,8	43,6	42,3	52,0	60,0
Bw	-15,1	**14,9**	42,8	33,4	36,6	39,2	34,6	44,9	49,8
CoG	-17,0	-9,0	**18,9**	32,9	35,9	34,6	36,7	48,4	54,1
Bal1	-10,6	-5,3	-0,9	**13,2**	29,5	35,0	33,2	43,5	48,0
Bal2	-7,2	0,9	5,6	6,3	**22,6**	36,2	37,4	43,1	48,8
Bal3	-10,4	-3,0	5,4	-0,7	7,6	**21,2**	36,1	43,9	48,7
MinGCI	-11,6	-0,9	0,9	-1,3	3,9	3,8	**18,7**	41,1	40,9
Naq	-15,6	-5,6	-5,1	-6,0	3,9	1,7	2,1	**24,4**	52,8
Qoq	-15,3	-2,0	-2,5	-2,0	6,5	5,3	10,6	4,4	**32,8**

IAIF	Fg	Bw	CoG	Bal1	Bal2	Bal3	MinGCI	Naq	Qoq
Fg	**25,9**	52,6	42,1	43,3	38,9	40,5	37,2	49,6	60,3
Bw	-10,7	**16,0**	36,4	36,6	30,7	33,3	28,5	44,2	52,9
CoG	7,8	3,6	**24,0**	30,0	26,7	25,1	32,3	41,6	48,9
Bal1	1,2	-2,1	12,5	**18,5**	29,1	27,7	26,6	40,1	48,0
Bal2	0,2	-1,5	10,5	2,6	**13,2**	24,4	22,0	34,4	39,8
Bal3	5,5	2,7	18,9	10,9	8,8	**20,1**	28,0	39,1	46,2
MinGCI	-3,7	-4,9	-0,8	-0,5	-1,3	-0,3	**7,5**	24,3	32,9
Naq	-2,1	-6,6	3,9	0,0	0,3	2,5	4,8	**21,6**	46,1
Qoq	-9,7	-12,2	-0,2	-4,8	-2,0	-1,4	-0,6	0,1	**24,7**

CPIF	Fg	Bw	CoG	Bal1	Bal2	Bal3	MinGCI	Naq	Qoq
Fg	**18,9**	49,9	35,2	34,7	33,7	31,5	27,6	45,5	63,8
Bw	-16,6	**14,4**	28,5	27,5	26,5	25,7	21,9	38,6	58,9
CoG	-8,4	-6,2	**7,9**	17,3	17,5	21,6	17,2	25,2	46,4
Bal1	-13,0	-10,3	-6,7	**2,7**	14,2	12,9	11,8	27,1	48,7
Bal2	-11,9	-9,2	-6,9	-8,7	**2,8**	12,1	8,2	25,0	45,2
Bal3	-9,3	-8,0	-10,5	-7,0	-6,1	**3,2**	12,7	21,9	44,2
MinGCI	-4,9	-3,6	-5,5	-5,2	-1,6	-5,6	**3,8**	20,7	36,7
Naq	-13,4	-11,0	-4,1	-11,2	-9,0	-5,5	-3,7	**13,2**	46,0
Qoq	-13,4	-13,0	-7,1	-14,5	-10,9	-9,5	-1,4	-1,4	**31,5**

Fig.2. Mutual information-based measures for the proposed features. *On the diagonal*: the relative intrinsic information. *In the bottom-left part*: the relative redundancy between two considered features. *In the top-right*: the relative joint information of the two considered features

REFERENCES

[1] P. Gomez-Vila, R. Fernandez, V. Rodellar, V. Nieto, A. Alvarez, R. Mazaira, and J. L. Godino, "Glottal source biometrical signature for voice pathology detection," *Speech Comm*.., vol. 51, pp. 759-781, 2008.

[2] M. Airas, and P. Alku, "Comparison of multiple voice source parameters in different phonation types," *Proc. of Interspeech 07*, pp. 1410-1413, 2007.

[3] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "A method for glottal formant frequency estimation," *Proc. of ICSLP 2004*, 2004.

[4] T. Drugman, T. Dubuisson, and T. Dutoit, "On the mutual information between source and filter contributions for voice pathology detection," *Proc. of Interspeech 09*, 2009.

[5] P. Alku, "Glottal wave analysis with pitch-synchronous iterative adaptative inverse filtering," *Speech Comm*., vol. 11, no 2-3, pp. 109-118, 1992.

[6] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation," *Proc. of Interspeech 09*, 2009.

[7] D. Veeneman, and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. on Signal Proc*., vol. 33, pp. 369-377, 1985.

[8] B. Bozkurt, and T. Dutoit, "Mixed-phase signal modeling and formant estimation using differential phase spectrums," *Proc. of VOQUAL' 03*, pp. 21-24, 2003.

[9] TKK Aparat website, http://aparat.sourceforge.net.

[10] K. Sjölander, and J. Beskow, "Wavesufer – an open source speech tool," *Proc. of ICSLP 2000*, vol. 4, pp. 464-467, 2000.

[11] T. Drugman, and T. Dutoit, "Glottal Closure and Opening Instant Detection from Speech Signals,"*Proc. of Interspeech 09*, 2009.

[12] J.B. Alonso, J. de Leon, I. Alonso, and A.M. Ferrer, "Automatic detection of pathologies in the voice by HOS based paramerers," *EURASIP Journal on Applied Signal Processing*, 2001:4, pp. 275-284, 2001.

[13] L. Huan, and H. Motoda, *Feature selection for knowledge discovery and data mining*, The Springer International Series in Engineering and Computer Science, vol. 454, 1998.

[14] T. Cover, and J. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, New York, 1991.

[15] Kay Elemetrics Corp. "Disordered Voice Database Model (version 1.03)", Massachusetts Eye and Ear Infirmary Voice and Speech Lab, 1994.